Face Detection on Distorted Images using Perceptual Quality–aware Features

Suriya Gunasekar[1], Joydeep Ghosh[2], and Alan C. Bovik[3]

Email: suriya@utexas.edu[1], ghosh@ece.utexas.edu[2], and bovik@ece.utexas.edu[3]

The University of Texas at Austin

ABSTRACT

We quantify the degradation in performance of a popular and effective face detector when human–perceived image quality is degraded by distortions due to additive white gaussian noise, gaussian blur or JPEG compression. It is observed that, within a certain range of perceived image quality, a modest increase in image quality can drastically improve face detection performance. These results can be used to guide resource or bandwidth allocation in a communication/delivery system that is associated with face detection tasks. A new face detector based on *QualHOG* features is also proposed that augments face–indicative *HOG* features with perceptual quality–aware *spatial Natural Scene Statistics (NSS)* features, yielding improved tolerance against image distortions. The new detector provides statistically significant improvements over a strong baseline on a large database of face images representing a wide range of distortions. To facilitate this study, we created a new *Distorted Face Database*, containing face and non–face patches from images impaired by a variety of common distortion types and levels. This new dataset is available for download and further experimentation at `www.ideal.ece.utexas.edu/~suriya/DFD/`.

1. INTRODUCTION

The advent of affordable digital storage devices and powerful, network pervasive visual data sharing websites such as Flickr, Facebook, Instagram etc., has caused an explosion of visual data that is being generated and shared at an exponentially growing rate. While the principal consumers of visual data are humans, practical machine vision deployments are becoming more common place. In both realms, automated methods for culling, sharing, organizing and understanding large volumes of visual content is highly desirable. Computer vision algorithms that aim at understanding visual content are being increasingly employed in real life applications such as image search, automated surveillance, human computer interfaces, etc. A primary component of many computer vision algorithms is some form of an object detection/recognition system. Such systems are often prone to performance degradation when the quality of the input images deteriorates. One such task that has resulted in successful commercial embodiments is the automatic face detection. Since face detection is usually a precursor to advanced tasks of recognition, expression tracking, etc., understanding the relationship between face quality and detectability is important.

There has recently been high interest in the development of automated image quality algorithms (IQA) that aim to accurately predict end–user quality–of–experience. These include *Full Reference (FR)* algorithms,[20,35] in which the fidelity of a test image to a presumed undistorted reference version is evaluated, *No Reference (NR)* algorithms,[24,26,31,37] which do not use any information from reference images, and the intermediate *Reduced Reference (RR)* algorithms,[7,21] which use partial information available about reference images. Among these, NR algorithms have the greatest potential for many practical settings, since references are seldom available. General purpose NR framework that rely on models of natural statistics of images have been recently shown to provide state–of–the–art performance in predicting perceived image quality.[25,26,31]

An exciting direction of inquiry is the interaction between visual quality and visual tasking. A small body of work exists on how quality affects biometric tasks (iris, face, fingerprint).[2,8,18,19] These papers study various image factors that affect the detection or recognition performance. For example, ISO/IEC 19794-5[1] specifies a list of factors such as spectacles, pose, centering, occlusion, expression, head shape, etc., that affect "face quality". While these do affect detection and recognition, there is no clear distinction between scene–dependent challenges like occlusion, illumination, etc., and the challenges imposed by traditional notions of "quality impairments" from capture, compression, processing, transmission, etc. In this paper, we are concerned with the later interpretation of "quality" as it affects the face detection performance. This is an important line of work as in many facial communication channels, the effect of such quality

impairments on detection/recognition can often be mitigated, e.g., by reallocating resources such as bandwidth. Further, in spite of numerous algorithms proposed in the past few decades,[17,38] face detection methods that are robust to image distortions have not been widely explored.

The IQA models described above are designed to predict the perceptual quality of digital images but have not been applied to visual task models involving faces. We explore the important question of whether the perceptual quality of facial images is a good predictor of the success of algorithm performance on visual tasks. The question is quite relevant given that the human visual apparatus is remarkably well adapted to analyzing faces. Early work by Rouse et. al. in this direction[27–29] show that perceptual FR IQA algorithms, including VIF and SSIM, correlate strongly with "recognizing thresholds" of human observers. However, the effects of quality on machine vision algorithms have not been rigorously evaluated and moreover FR algorithms are of limited use in this regard.

We begin by investigating the effects of blindly measured NR image quality degradations on face detection performance. A trade–off exists between ground truth image distortion and face detection performance. In many image/video communication channels, the distortion levels could be adjusted using channel parameters to obtain a required level of face detection performance. However, in most scenarios accurate measures of distortion types and levels is not available. We then show that as with true distortion levels, over a range of objective quality scores delivered by a high–performance NR image quality model, moderate improvements in predicted quality can significantly aid face detection performance.

Secondly, we show that the use of easily computable "quality–aware" *spatial Natural Scene Statistics (NSS)* features[24] has the potential to greatly assist the design of more robust face detection algorithms. The widely–used Histogram of Oriented Gradients (HOG) based face detection algorithm[10] is used as the baseline in our experiments. We use this model because it is flexible and easily reconfigured to enable the inclusion of features related to image quality.

Finally, existing face detection datasets* consist of samples of face and non–face patches. However, our goal is to investigate the performance of face detectors on images corrupted by common distortions such as gaussian blur and JPEG compression. Such distortions are usually global, so distortions isolated on local patches may be expected to exert a different effect on detection performance as compared to distortion of the entire image. Thus, we curated a new Distorted Face Database (DFD), from the web for our experiments. This new dataset consists of face and non–face patches from images that were distorted with known distortion types and levels. The dataset is available for download and further experimentation at `www.ideal.ece.utexas.edu/~suriya/DFD/`.

The main contributions we make here are as follows:

1. The performance degradation of a widely used generic face detector with respect to the response of an NR image quality algorithm called NIQE is studied on images distorted by three common distortions: additive white gaussian noise, gaussian blur and JPEG compression. We experimentally show that over a certain range of NIQE scores, a modest improvement in objective image quality can significantly improve face detection performance.

2. A new face detector based on *QualHOG* features is proposed that augments face–indicative *HOG* features with perceptual quality–aware *spatial NSS* features, thus supplying improved tolerance against distortion. We experimentally quantify the degree of resulting improvements.

3. A new Distorted Face Database (DFD), was created that has face and non–face patches from images that were distorted using known distortion types and levels.

The rest of the paper is organized as follows. In Section 2, we review relevant literature on image quality assessment and face detection algorithms. The distortion types investigated and the proposed model for robust face detection are discussed in Section 3. In Sections 4 and 5 we describe the experimental setup and the results, respectively. We conclude with directions for future work in Section 6.

*http://www.facedetection.com/facedetection/datasets.htm

2. RELATED WORK

In this paper we primarily borrow ideas from two basic problems in vision science and computer vision: image quality assessment and face detection. *Image quality assessment (IQA)* aims at predicting the quality of a given image as perceived by human users. The performance of IQA models are assessed by measures of correlation between objective predicted quality scores and aggregated human opinions (Differential Mean Opinion Score (DMOS)) on a set of representative test images. *Face detection* is a fundamental problem in various computer vision applications including camera focusing. Efficient and accurate algorithms for face detection have been widely developed over the past few decades. The problem of face detection involves accurately identifying the region(s) in an arbitrary image that corresponds to human face(s). In the rest of this section, we review some relevant literature pertaining to these two problems.

As stated previously, IQA algorithms can be broadly categorized as *Full Reference (FR)*, *Reduced Reference (RR)* and *No Reference (NR)* models. While the presence of a reference image or information regarding the reference simplifies the problem, in real–life applications FR and RR algorithms are limited in scope as the reference information is generally unavailable at nodes where quality computation is undertaken. Hence, we concentrate only on NR IQA models as they are much more likely to be of use in practical vision applications. Early NR IQA models were distortion specific.[4,39] Such algorithms extract distortion–specific features that relate to loss of visual quality, such as ringing, blur, edge–strength at block boundaries, etc. While these provide satisfactory performance for specific distortion types, often, the distortion type that is actually encountered is unknown beforehand or is poorly modeled. Thus, a few distortion–independent approaches to NR IQA problem have been proposed recently.[24,26,31] These models are based on the hypothesis that natural images follow regular statistical properties that are modified by the presence of distortions. Deviations from the regularity of these *natural scene statistics (NSS)* are indicative of the perceptual quality of images. Hence, models based on the quantification of *naturalness* of an image are capable of providing distortion–independent measure of perceived image quality.

For example, the *DIIVINE* index[26] deploys summary statistics derived from the NSS models of wavelet coefficients. These features are used to first identify most likely distortion types followed by distortion specific quality assessment. A similar approach named *BLIINDS–II*,[31] operates in the DCT domain. A small number of features are computed from an NSS model of block DCT coefficients. These features are in turn used to train a regression model that delivers accurate quality predictions. While both DIIVINE and BLIINDS–II deliver superior performance for assessing image quality, computation of the features involved is expensive and hence deploying these models in real time is difficult.

Scalable transform–free (spatial) models for NR IQA were recently developed by Mittal et al.[24,25] The *BRISQUE* and *NIQE* indices proposed in these works operate directly on multiscale spatial pixel data and hence are inexpensive to compute. These models are based on the statistics of locally debiased and divisive normalized luminance coefficients that quantify the deviation from *naturalness* of an image due to the presence of distortions. BRISQUE uses quality–aware spatial features to train a regression model for IQA, while NIQE develops a model for undistorted "pristine" images and measures deviations of the statistics of the test image from the pristine image model. Despite using purely spatial features, these models show performance comparable to DIIVINE and BLIINDS–II at a small fraction of the required computation. Going forward, we will use the Spatial NSS features used by BRISQUE and NIQE as quality–aware features.

Some of the early work on face detection has been surveyed by Hjelmas et al., Yang et al.,[17,36] and more recently by Zhang et al.[38] Early work in face detection has been categorized as *knowledge based methods*, which use predefined rules to detect faces, or as *feature invariant methods*, which use pose and lighting invariant features, or as *template matching methods*, which detect faces by matching against pre–stored representative face images, or finally as *appearance based methods*, which model faces from a set of representative training faces.

Most of the recent algorithms for face detection could be categorized as appearance based methods. A typical practice is to collect certain indicative features from a training set of face and non–face image patches and use machine learning algorithms to learn a classifier for detecting other faces. The two key variants among these approaches are the type of features used and the kind of classifier employed in learning the models.

The Viola–Jones algorithm[33] for face detection has had a large impact on face detection research because of its low computational complexity that has made face detection feasible in real time. The algorithm uses simple Harr–like features to train weak classifiers in a multi–stage boosting algorithm.[23] Boosting algorithms have been a popular choice in the literature. AdaBoost, RealBoost, and GentleBoost are some of the popular methodologies in this framework and they have been compared by Lienhart et al. and Brubaker et al.[6,22]

More recently, regional image statistics features are being used increasingly for face detection. With the advent of more complex features, various single stage classifiers like Bayesian classifiers and support vector machines (SVMs) have gained popularity. For example, Dalal et al.,[10] introduced a popular regional statistics based feature engine called the *Histogram of Oriented Gradients (HOG)* and used a linear SVM classifier to detect humans in an image. These features are invariant to 2D rotations and illuminations. An extensive survey of the various other features used by recent face detection algorithms is provided by Zhang et al.[38] The baseline used for comparison in this paper is an adaptation of the human detector proposed by Dalal et al.,[10] for the problem of face detection.[3]

Studying the effects of quality impairments on detection and recognition tasks is of interest as it can be exploited to mitigate the effects of such impairments on relevant tasks. Some work can be found in the literature that study the effect of perceived image quality on object detection/recognition performance.[15,27–29] Rouse et al. take a broad view of quality vs. tasking.[27–29] Recognizing the importance of perceptual principles in both visual tasks and in quality assessment, the authors study human "recognition thresholds" of objects as a function of objective image quality as measured by the FR algorithms multiscale SSIM and VIF. They find that perception–driven FR IQA indices can indeed successfully predict image recognizability.[27,28] Likewise, Gala et al.[15] find that the SSIM metric can be used to predict the performance of tracking algorithms with a high degree of confidence. However, as mentioned previously, FR and RR algorithms are limited in their applicability and hence we investigate face detection performance as a function of image quality predicted by a state–of–the–art NR algorithm. This is of particular importance in forthcoming wireless vision applications, where intelligent, robust blind algorithms are needed, where severe distortions occur, and where facial images are becoming increasingly important in both consumer and/or security applications.

3. QualHOG BASED FACE DETECTOR

In this section we first describe the types of image distortions that we consider. A new quality–aware face detector called QualHOG, which uses face–indicative HOG features and quality–indicative spatial NSS features is then described.

3.1 Image Distortions

We consider three basic types of distortions that commonly occur in digital devices and over communication channels. The image is denoted by a matrix I, such that $I(i,j)$ represent the $(i,j)^{\text{th}}$ pixel in the image I.

1. **AWGN(σ_N^2), Additive White Gaussian Noise**: This is a local distortion, in which a zero mean gaussian noise of variance σ_N^2 is added independently to each pixel. This distortion is parameterized by the variance, σ_N^2.

$$\tilde{I}(i,j) = I(i,j) + N_{ij}, \text{ such that } N_{ij} \sim \mathcal{N}(0, \sigma_N^2) \tag{1}$$

   where $\mathcal{N}(\mu, \sigma^2)$ is a gaussian distribution with mean μ and variance σ^2. This is a common model for a broadband device or channel noise.

2. **Gblur(σ_B), Gaussian Blur**: This is a global distortion in which each pixel is blurred through convolution with a gaussian low pass filter of standard deviation σ_B. For computational ease the gaussian kernel is truncated at $6\sigma_B$. The discrete truncated gaussian filter in two dimensions is given as follows:

$$G(x,y) = \frac{1}{2\pi\sigma_B^2} e^{-\frac{x^2+y^2}{2\sigma_B^2}} \tag{2}$$

   where $-\lceil 3\sigma_B \rceil \leq x \leq \lceil 3\sigma_B \rceil$ and $-\lceil 3\sigma_B \rceil \leq y \leq \lceil 3\sigma_B \rceil$. An image with gaussian blur is given by $\tilde{I} = I * G$

$$\tilde{I}(i,j) = \sum_{x=-\lceil 3\sigma_B \rceil}^{\lceil 3\sigma_B \rceil} \sum_{y=-\lceil 3\sigma_B \rceil}^{\lceil 3\sigma_B \rceil} I(i+x, j+y) G(x,y) \tag{3}$$

   This is a common model for lens blur.

3. **JPEG(Q), JPEG compression**: This is the most commonly used lossy compression method for digital photography. The trade–off between storage size and image fidelity is controlled by a "quality factor", $0 \leq Q \leq 100$, where $Q = 100$ corresponds to no compression while lower values of Q lead to higher compression and lower image quality. Note that while Q is generally monotonic with the perceived quality of a compressed image, it is a poor predictor of the perceptual image quality. This compression scheme first converts the spatial image into the frequency domain using a discrete cosine transform (DCT). In the DCT domain the DCT coefficients are quantized to reduce storage requirements. The degree of quantization is controlled by the Q factor. If G is the DCT matrix of image I, the quantized DCT matrix, $\tilde{G}$ is given by:

$$\tilde{G}(i,j) = \text{round}\left(\frac{G(i,j)}{\mathbb{Q}(i,j)}\right) \tag{4}$$

where the quantization matrix, $\mathbb{Q}$ (dependent on Q), which is of the same size as G, is designed to provide higher resolution in frequencies that are hypothesized to be perceptually more important.

3.2 QualHOG face detector

The QualHOG patch descriptor consists of the following components:

1. **Spatial NSS**: The spatial NSS features used in QualHOG were proposed Mittal et al.[24] to acomplish blind quality assessment and consist of parameters describing the natural scene statistics of spatial components. The image patch, I, is preprocessed using local mean removal and divisive normalization:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{5}$$

where (i,j) are spatial indices, $\mu(i,j)$ and $\sigma(i,j)$ are the mean and variance, respectively, of neighborhood pixels weighted by a truncated symmetric 2–D gaussian.

The motivation for these NSS features lies in statistical models of photographs and in perceptual models of human vision. It is well established that early stages of human vision process images locally. These processes have evolved to encode images using natural statistics for optimal neural transmission.[11,14] Ruderman[30] hypothesized that the neural channel for transmitting visual signals were constrained by the variance of the signals and hence the optimal coding of images could be attained using gaussian statistics. He established that local mean subtracted and divisive normalized pixel values of natural images (as in Equation 5) have gaussian histograms. The mean subtraction in the numerator of the equation is a center–surround band pass operation that approximates post–retinal ganglion processing to obtain residual images with lower entropy; apparently to accomplish predictive coding.[32] The divisive normalization by σ in the denominator models the adaptive gain control process (AGC) in the visual cortex that accomplishes contrast masking as a byproduct.[16,34] The constant C, also known as the saturation constant, stabilizes the division.

A white gaussian model of (5) is quite regular across good–quality photographic images. However, when images are distorted, histograms of pixels after preprocessing using (5), are generally no longer gaussian. Extensive experimentation with IQA models has shown that the distorted image histograms subject to (5) can be fit using a generalized gaussian distribution (GGD) and the deviation of an image from the "true naturalness" quantifies the distortion types and levels. A zero mean GGD is parameterized by (α, β) and is given as:

$$f(x;\alpha,\beta) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} \exp\left(\frac{-|x|}{\beta}\right)^{\alpha} \tag{6}$$

Moreover, it has been observed that distortions typically introduce unnatural spatial dependencies, which can be measured by examining the distributions of local image correlations.[5] A set of directional (horizontal,vertical and diagonal) spatial features are computed as:

$$\begin{aligned} H(i,j) &= \hat{I}(i,j)\hat{I}(i,j+1) \\ V(i,j) &= \hat{I}(i,j)\hat{I}(i+1,j) \\ D1(i,j) &= \hat{I}(i,j)\hat{I}(i+1,j+1) \\ D2(i,j) &= \hat{I}(i,j)\hat{I}(i+1,j-1) \end{aligned} \tag{7}$$

The histograms of each directional components, $\{H(i,j)\}$, $\{V(i,j)\}$, $\{D1(i,j)\}$ and $\{D2(i,j)\}$ are fit using a zero mode asymmetric generalized gaussian distribution (AGGD), which is parameterized by $(\gamma, \beta_l, \beta_r)$ as:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\gamma}{2(\beta_l+\beta_r)\Gamma(1/\gamma)} \exp\left(\frac{-|x|}{\beta_l}\right)^{\alpha}, \text{if } x \leq 0 \\ \frac{\gamma}{2(\beta_l+\beta_r)\Gamma(1/\gamma)} \exp\left(\frac{-|x|}{\beta_r}\right)^{\alpha}, \text{if } x > 0 \end{cases} \tag{8}$$

Finally, the statistical mean of each AGGD fit is computed as:

$$\eta = (\beta_r - \beta_l)\frac{\Gamma(2/\gamma)}{\Gamma(1/\gamma)} \tag{9}$$

Using, the $(\gamma, \beta_l, \beta_r, \eta)$ estimates along the four directions and the (α, β) estimates of the GGD fit to the histogram of $\{\hat{I}(i,j)\}$, $18D$ features are computed at two scales leading to a $36D$ *spatial NSS* feature vector.

2. **HOG**: The HOG descriptor was first introduced by Dalal et al.[10] It is a widely used feature descriptor for various object detection tasks.[13] To compute the HOG features, a detection window is divided into dense overlapping blocks of size 16×16 with a stride of 8×8 pixels. Each block is further divided into 2×2 cells and a histogram of gradients in 9 orientations is computed within each cell. All the histograms within a patch are concatenated to form the HOG feature descriptor.

   This feature descriptor quantifies the gradient structure within a block which characterizes local edge directions. The appearance of an object in a detection window can be largely captured by the edge directions within indexed blocks. The local intensities are initially contrast normalized (before computation of the gradients) to provide illumination invariance. Thus, a discriminative classifier trained on histograms of oriented gradients extracted from dense set of local blocks in a detection window is capable of generalizing to other objects.

QualHOG: The quality aware *QualHOG* descriptor is obtained by simply concatenating the HOG and spatial NSS features. In our experiments, the detection windows are of size 80×64, which gives a HOG feature vector of length 2268, which combined with the $36D$ Spatial NSS features yields 2304 dimensional QualHOG feature vector. The motivation behind appending perceptually relevant quality–aware features to conventional object detection features is that the optimal decision boundary in the HOG vector space varies non–trivially as a function of input image/video quality. By appending spatial NSS features to the HOG feature vector and passing this to a linear SVM, we effectively model a quality dependent boundary shift in the space spanned by the HOG features.

Fast Spatial NSS: Since QualHOG is intended to be used in a scanning window approach, we first implemented a fast computation algorithm using integral images to allow efficient Spatial NSS feature computation within rectangular windows in an image. By using this *Fast Spatial NSS* implementation, it is only necessary to first compute the integral images at each scale in an image pyramid. Computation of spatial NSS features for any rectangular window is near–instantaneous thereafter.

QualHOG based face detector: Linear support vector machines (SVMs)[9] were trained using the QualHOG features from face and non–face patches. Specifically, we use a soft–margin SVM with a slack penalty that simultaneously maximizes the margin while minimizing the training error. SVMs with non–linear kernels were also tried in the initial experiments, however, they require much longer computational time and did not provide any significant improvement in the results.

4. EXPERIMENTS

As discussed in Section 3, the distortions considered in this paper (except for additive white noise) are global and hence we cannot use existing face databases that have only face and non-face patches. Instead we require full images which could be distorted by known distortion types and levels. Thus, as a first step, we created a new Distorted Face Database (DFD) of facial images from images available freely on the internet. To keep the task simple, we chose images with mainly frontal faces. A total of 215 images were crawled, each with one or more frontal faces. These images were manually ensured to be

of high quality with no visible distortions. This set of 215 images was divided into 150 training images and 65 test images. The faces in these images were manually annotated. A total of 1231 faces were present in the training set of images and 393 were present in the test set.

For simplicity, we demonstrate our model only at one scale and hence we designed a system that detects faces in patches of size 80×64. In order to obtain training and testing face samples of the required dimensions, we resized the images so that the average sizes of the faces within an image are 80×64. Also, in the image selection process, care was taken to ensure that in case of multi–face images, the sizes of the faces were not widely different. In this way, on the resized images, a 80×64 sized bounding box centered at the faces captures the facial content reasonably well.

Next, the images were modified in various ways to create distortions. The following distortion types were introduced at different levels on the training and test datasets.

1. AWGN: The *imnoise()* function in MATLAB was used to introduce additive white gaussian noise to the images. 10 levels of AWGN were added to the images with the noise variance parameters varying over a log scale, $\sigma_N^2 = \{4.5 \times 10^{-5}, 0.0001, 0.0003, 0.0009, 0.0025, 0.0065, 0.02, 0.05, 0.15, 0.36\}$.
2. GBlur: The *imfilter()* function in MATLAB was used to introduce gaussian blur at 10 levels. The standard deviation of the gaussian filter was sampled over a log scale, $\sigma_B = \{0.4, 1.0, 2.3, 3.6, 4.5, 6.0, 7.4, 12.0, 20.0, 32.0\}$.
3. JPEG: The *imwrite()* function in MATLAB was used to produce JPEG compressed images at 10 levels of distortion. The Q factor controlling the quality of the image was also sampled on a log scale, $Q = \{90, 60, 40, 25, 15, 10, 7.5, 5.0, 3.0, 2.0\}$.

Training the face detector From each of the above sets of training images, the 1231 manually annotated faces were cut out to provide positive samples for each dataset. A random subset set of 1500 negative patches were initially selected from the non–face parts of the images in each training dataset.

Soft–margin linear SVMs were first trained using QualHOG features extracted on the positive and negative samples from different combinations of the training datasets described above. As a baseline, analogous classifiers were trained using just the HOG features. Hereafter, we follow the following terminology. A face detector trained on QualHOG and HOG features of samples from pristine images alone will be called *QualHOG–Prist* and *HOG–Prist* respectively. Similarly, face detectors trained on QualHOG and HOG features of samples from pristine images and images from $L1$ to Ln levels of distortion of distortion type D, are denoted as *QualHOG–D–L1-n* and *HOG–D–L1-n* respectively (for example *QualHOG–AWGN–L1-4* refers to the face detector trained on QualHOG features of samples from pristine training images and images distorted with AWGN of variances $\{4.5 \times 10^{-5}, 0.0001, 0.0003, 0.0009\}$).

The implementation of soft-margin linear SVM from LIBLINEAR[12] was used in the experiments. For each classifier, the non–face regions of the 150 training images were searched exhaustively for false positives, also referred to as "hard negatives". A maximum of 1000 hard negatives were obtained for each training dataset. The classifiers were then retrained using the augmented set of negative samples (the initial 1500 negative samples + hard negatives) to produce the final detector. This retraining process is adapted from the work by Dalal et al.,[10] where the authors observed a significant improvement in the performance of each detector. We observed similar behavior in cross–validation performance in our experiments. Finally, for each type of face detector (QualHOG and HOG based), the parameter C for the soft–margin SVM was chosen via cross–validation by doing a grid search on the log scale.

Testing As mentioned earlier, the 393 faces annotated on each of the test datasets were cut out to obtain positive test samples and an exhaustive set of 17494 negative samples were extracted from the non–face parts of the test images from the corresponding dataset. The *area under precision recall curve (AUPR)* was used as the evaluation metric since the test dataset is highly skewed as compared to the training dataset. Typically, the continuous output of a classifier is thresholded to determine the discrimination boundary. Precision is defined as the fraction of detected positives that are faces, i.e., the ratio of true positives to the detected positives. Recall is defined as the fraction of actual positives that is detected, i.e., the ratio of true positives to the total number of positives. Precision–recall curves for a system plot the trade–off between precision (y axis) and recall (x axis) as the discrimination threshold is varied.

Allocating images to NIQE bins One of our contributions is to study the degradation of face detection performance as a function of automatically reported objective NR image quality scores. We used the NIQE index[25] as a surrogate for perceived image quality because of its excellent cost–performance. In order to evaluate the performance of face detectors against NIQE levels, we require test datasets at different NIQE levels. It is not easy to create an image at a pre–defined NIQE level. Thus, in order to obtain test datasets at different NIQE levels, we first computed NIQE scores for all the distorted images, AWGN–L1–10, GBlur–L1–10 and JPEG–L1–10, in the test dataset. The images within each distortion type, AWGN, GBlur and JPEG, were in turn binned into 10 datasets, with the mean NIQE score within each bin serving as the representative NIQE score for that bin.

5. RESULTS

In practical settings, precise information regarding the distortion types and distortion levels afflicting an image are difficult to estimate. The NIQE image quality index, described in Section 2, on the other hand, is a high performance distortion agnostic algorithm that does not rely on any form of distortion models. Further, the Spatial NSS features used to compute NIQE scores are computationally inexpensive as compared to other NR quality scores.[26,31] We therefore use NIQE scores as surrogates for perceptual distortion levels. However, as a sanity check, we first assessed the NIQE scores of images against all of the various distortions and over each range of levels. The NIQE scores of images distorted by AWGN, gaussian Blur and JPEG distortions are shown in Figures 1, 2 and 3 respectively. As expected, a strong positive correlation between degree of distortion and NIQE scores is observed for all distortion types.

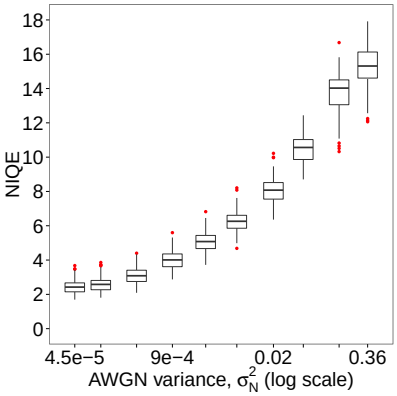

Figure 1: NIQE vs AWGN

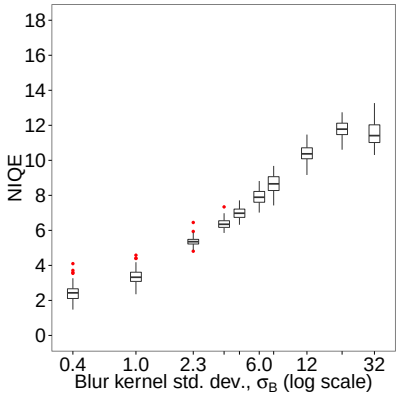

Figure 2: NIQE vs gaussian Blur

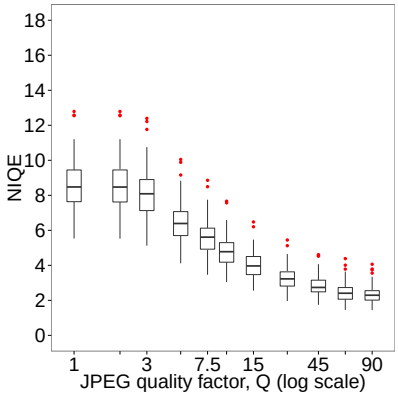

Figure 3: NIQE vs JPEG

We next studied the performance degradation of the baseline HOG based face detector, *HOG–Prist*, on distorted images that are quality assessed by NIQE. In order to evaluate the performance of face detectors against NIQE scores, we first binned the images from the test dataset images, which were distorted by multiple degrees of each distortion type, into 10 discrete NIQE levels, then evaluated the performance of the baseline HOG–Prist face detector in each bin. Figure 4 plots the performance degradation of the *HOG–Prist* face detector against NIQE for all distortion types.

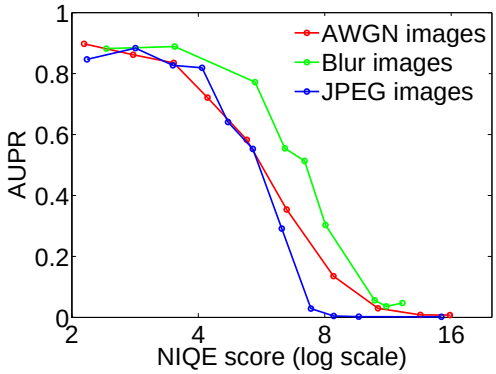

Figure 4: Performance degradation of *HOG–Prist* face detector with perceived quality measured as NIQE

It is not surprising that the degradation of face detection performance with increasing NIQE score is largely monotonic. It is, however, interesting to note that a region of steep decline of face detection performance exists for NIQE scores in the range 5–8, in which small improvements in image quality can yield significant enhancements to face detection performance. This trade–off between perceived image quality and face detection performance could be immediately exploited in the design of optimum communication channel parameters in a facial image communication system. Moreover, it is also interesting to note that for a given level of predicted image quality, the HOG–Prist face detector is slightly more tolerant of quality impairments due to gaussian blur than those due to other distortions.

5.1 QualHOG vs HOG

The second contribution that we make is a new QualHOG face detector which is more robust to image distortions as it combines the face indicative cues with the quality–aware image features. The performance of QualHOG and HOG based face detectors trained and tested on the new database is discussed in this section.

We first trained two face detectors, *QualHOG–Prist* and *HOG–Prist*, that use QualHOG and HOG features respectively, using samples from only pristine images. We then developed unified distortion–independent models, *QualHOG–All* and *HOG–All*, wherein training samples from all three distortion types were used to train the face detectors. Finally, for each distortion type, AWGN, GBlur and JPEG, we also trained distortion–dependent QualHOG and HOG based face detectors using samples with increasing levels of distortions, *QualHOG–[D]–L1*, *QualHOG–[D]–L1-2*, ..., *QualHOG–[D]–L1-10* and *HOG–[D]–L1*, ..., *HOG–[D]–L1-10*, respectively, where, *[D]* is a placeholder for each distortion type, AWGN, GBlur, JPEG etc.

5.1.1 Performance vs NIQE scores

As the knowledge of distortion type is often unavailable, we first show results of distortion–independent face detectors. We created distortion–unaware test datasets by binning the NIQE scores from all distorted images in the overall test dataset (all distortion types and all distortion levels) into 20 bins. We then plotted the 5–fold cross validation performance of the distortion–independent face detectors, *QualHOG–Prist*, *HOG–Prist*, *QualHOG–All* and *HOG–All* on test images in each bin. Note that there were a few outliers (possibly due to approximations in NIQE binning) and we used local linear smoothing to obtain the interpolated plots. These results along with cross–validation error bars are plotted in Figure 5. The NIQE scores are represented on a log–scale on the horizontal axis.

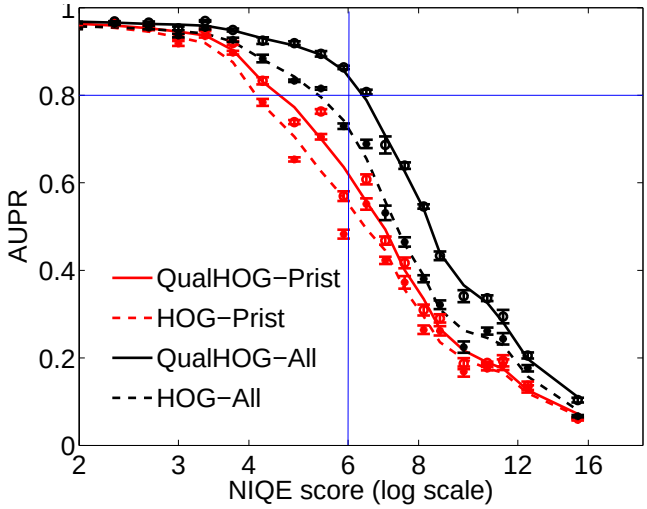

Figure 5: AUPR vs NIQE for distortion independent face detectors. For a performance level of 0.8, the QualHOG–All is tolerant to images with NIQE scores up to ~ 6.5 and HOG–All can handle images with NIQE score of up to ~ 5.

It can be seen that QualHOG based face detectors show significant improvements over the HOG based ones. For face detectors trained on just the pristine images, the improvement of QualHOG–Prist over HOG–Prist is small. This is because there is not much information about the distortions in the training samples that the spatial NSS features used in QualHOG could deliver a benefit from.

Training on distorted images improves the performance of both HOG and QualHOG based face detectors. The HOG based face detector is constrained to a single detection boundary in the HOG vector space to capture the discriminating characteristics across all distorted images. However, using the quality–aware spatial NSS features, QualHOG face detectors are capable of modeling a quality dependent boundary shift in HOG feature space. Thus, the improvement from training on distorted samples is higher for QualHOG. For a given level of target performance required, QualHOG–All uniformly provides higher tolerance for degradation in quality. For example, QualHOG–All can provide a reasonably good AUPR of 0.8 for images with NIQE scores less than ~ 6.5, while the performance of HOG–All rapidly falls below 0.8 if the test images with NIQE scores greater than ~ 5 are included. In the narrow range of NIQE scores where there is a steep decline in face detection performance of HOG–All, substantial gains were observed for QualHOG–All. For example, within a NIQE range centered at 6.0, QualHOG-All achieves an AUPR of 0.86 while the AUPR achievable by HOG–All is only 0.72. Thus, the QualHOG based face detectors are able to achieve acceptable face detection performance at much higher levels of visual impairment than what is currently possible. Finally, we note that the error bars were very small and we avoided reporting them in the rest of the paper.

For the performance analysis on individual distortion types, AWGN, gaussian blur and JPEG, face detectors trained on samples from specific distortion types were evaluated on test datasets from 10 discrete NIQE levels for each distortion type. To avoid clutter from reporting the performances of all the distortion–dependent face detectors, we report the results of only the best performing detector for each distortion type, along with the distortion–independent detectors, *QualHOG–Prist*, *HOG–Prist*, *QualHOG–All* and *HOG–All*. The best performing face detectors were separately chosen for the HOG and QualHOG based detectors. These results are compared in Figure 6, 7 and 8, for AWGN, GBlur, and JPEG distortions respectively. The best performing QualHOG and HOG face detectors for AWGN, gaussian blur and JPEG were, QualHOG–AWGN–L1-6, HOG–AWGN–L1-6, QualHOG–GBlur–L1-8, HOG–GBlur–L1-8, QualHOG–JPEG–L1-6 and HOG–JPEG–L1-8 respectively.

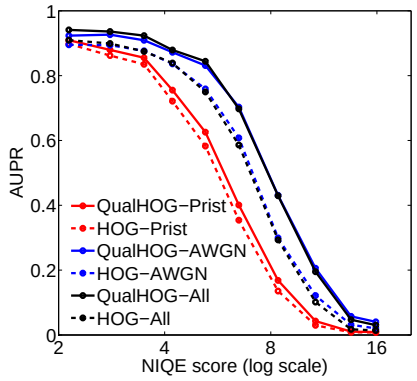

Figure 6: AUPR vs NIQE for images distorted with AWGN

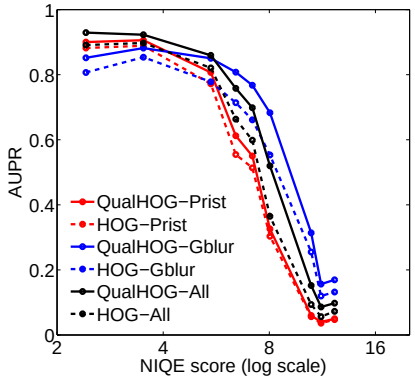

Figure 7: AUPR vs NIQE for images distorted with gaussian blur

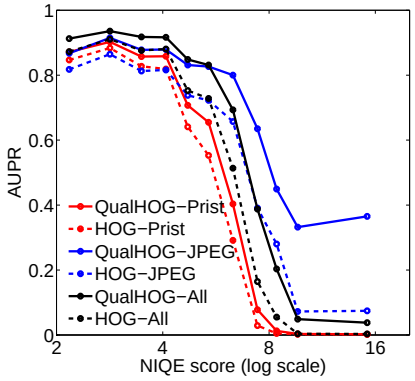

Figure 8: AUPR vs NIQE for images distorted with JPEG distortion

For all the distortions considered, QualHOG based detectors again provide higher tolerance against degraded images. The improvement is higher for AWGN and JPEG distortions, while the models show comparable performance on gaussian blur distortion. It was expected that face detectors trained on samples with specific distortion types improve upon the face detectors trained on only the pristine images when tested on individual distortion types. However, even the distortion–independent QualHOG–All face detector trained on all the distortion types and distortion levels was able to achieve robust performance comparable to the distortion–specific ones. These observations validate our hypothesis that quality–aware image features can aid in building distortion–robust face detectors.

5.1.2 Performance vs ground truth distortion levels

To complete our analysis, we also compare the performance of the face detectors against the ground truth distortion levels for the three types of distortions considered. The face detectors trained as above were tested on test datasets from all 10 levels of appropriate distortion types (with distortion levels as specified in Section 4). Again, to avoid clutter we report the results of only the best performing detector for each distortion type, along with *QualHOG–Prist*, *HOG–Prist*, *QualHOG–All* and *HOG–All*. These results are shown in Figure 9, 10 and 11, for AWGN, GBlur, and JPEG distortions, respectively. The distortions levels are represented on a horizontal log–scale.

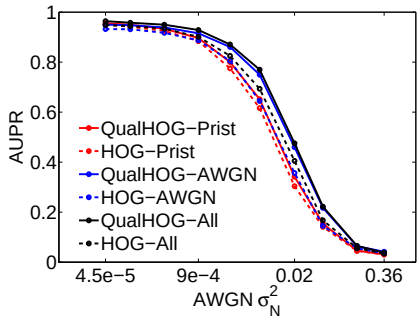

Figure 9: AUPR vs AWGN variance, σ_N^2

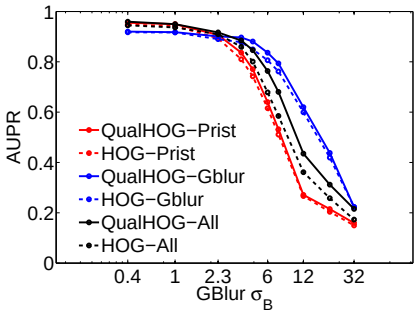

Figure 10: AUPR vs standard deviation of gaussian blur kernel, σ_B

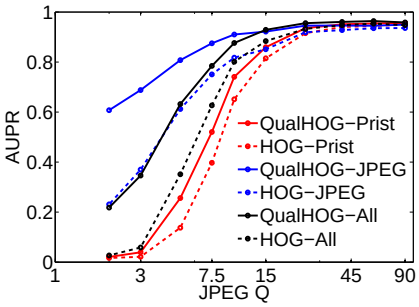

Figure 11: AUPR vs JPEG quality factor, Q

It can be seen that similar trends are observed for face detection performances measured at various ground truth distortion levels and those measured at discrete NIQE levels.

6. CONCLUSIONS

In this paper we first established that the easily computable NR image quality score, NIQE, is effective as a proxy for actual distortion levels when evaluating the trade–off between face detection performance against image impairments arising from three common distortion types, AWGN, gaussian blur and JPEG. The performance of generic HOG–based face detectors was found to degrade rapidly for NIQE scores greater than 4. It was also observed that for NIQE scores in the 5–8 range, a modest improvement in perceived image quality measures drastically improves face detection performance. This region can be fruitfully targeted when allocating resources in constrained settings. Another interesting observation was that, face detector performances are consistently more tolerant of quality impairments due to gaussian blur than those due to other distortions considered.

Secondly, we showed that QualHOG features, which combine face indicative HOG features with quality–aware spatial NSS features are more effective at learning a face detector that is robust to common and important image distortions. The QualHOG based face detectors show significant improvement over their HOG based analogues when trained on distorted images. In a practical distortion–unaware setting, the QualHOG–All face detector typically produced reliable results ($AUPR \geq 0.8$) for test datasets with NIQE scores of up to 6.5, while HOG–All provided equivalent performance on images with NIQE score up to 5. Interestingly, in spite of being distortion–independent, QualHOG–All also provides comparable performance to distortion–specific QualHOG models, when tested on individual distortion types. Thus, the QualHOG based face detectors are able to achieve acceptable face detection performance at much higher levels of visual impairments than what is currently possible.

Acknowledgement

This research was funded by the NSF grant IIS-1116656.

REFERENCES

[1] ISO/IEC 19794-5. Information technology - biometric data interchange formats - part 5: Face image data, 2005.

[2] M. Abdel-Mottaleb and M. H. Mahoor. Application notes - algorithms for assessing the quality of facial images. *Computational Intelligence Magazine, IEEE*, 2007.

[3] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol. Face recognition using hog – ebgm. *Pattern Recognition Letters*, pages 1537 – 1543, 2008.

[4] R. V. Babu, S. Suresh, and P. L. Perkis. No-reference jpeg image quality assessment using gap-rbf. *Signal Processing*, 87(6):1493–1503, June 2007.

[5] A. C. Bovik. Automatic prediction of perceptual image and video quality. To Appear, 2012.

[6] S. C. Brubaker, J. Wu, J. Sun, M. D. Mullin, and J. M. Rehg. On the design of cascades of boosted ensembles for face detection. *International Journal on Computer Vision*, pages 65–86, 2008.

[7] P. Campisi, M. Carli, G. Giunta, and A. Neri. Blind quality assessment system for multimedia communications using tracing watermarking. *Signal Processing, IEEE Transactions on*, 51, 2003.

[8] Y. Chen, S. C. Dass, and A. K. Jain. Localized iris image quality using 2-d wavelets. In *Proceedings of the 2006 international conference on Advances in Biometrics*, 2006.

[9] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3), September 1995.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005.

[11] J. E. Dowling. *The Retina: An Approachable Part of the Brain*. Belknap Press of Harvard University Press, 1987.

[12] 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*, 9:1871–1874, 2008.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 1627–1645, 2010.

[14] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, pages 2379–2394, 1987.

[15] A. Gala and S. Shah. Joint modeling of algorithm behavior and image quality for algorithm performance prediction. In *Proceedings of the British Machine Vision Conference*, 2010.

[16] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Vis Neurosci*, 1992.

[17] E. Hjelmas and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, pages 236 – 274, 2001.

[18] R. L. V. Hsu, J. Shah, and B. Martin. Quality assessment of facial images. In *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the*, 2006.

[19] N. D. Kalka, V. Dorairaj, Y. N. Shah, N. A. Schmid, and B. Cukic. Image quality assessment for iris biometric. In *SPIE 6202: Biometric Technology for Human Identification III*, pages 445–452. Springer, 2002.

[20] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for image artifacts based on human visual sensitivity. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, 1994.

[21] Q. Li and Z. Wang. Reduced-reference image quality assessment using divisive-normalization-based image representation. *IEEE J. Selected Topics in Signal Processing*, 2009.

[22] R. Lienhart, E. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM 25th Pattern Recognition Symposium*, pages 297–304, 2003.

[23] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *International Conference on Image Processing*, 2002.

[24] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *Image Processing, IEEE Transactions on*, pages 4695–4708, 2012.

[25] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *Signal Processing Letters, IEEE*, pages 209–212, 2013.

[26] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *Image Processing, IEEE Transactions on*, pages 3350–3364, 2011.

[27] D. M. Rouse and S. S. Hemami. Quantifying the use of structure in cognitive tasks. *SPIE Conf. Human Vision Electronic Imaging*, 2007.

[28] D. M. Rouse and S. S. Hemami. Analyzing the role of visual structure in the recognition of natural image content with multi-scale ssim. *SPIE Conf. Human Vision Electronic Imaging*, 2008.

[29] D. M. Rouse, R. Ppion, S. S. Hemami, and P. L. Callet. Image utility assessment and a relationship with image quality assessment. In *Storage and Retrieval for Image and Video Databases*, 2009.

[30] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, pages 517–548, 1994.

[31] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *Image Processing, IEEE Transactions on*, pages 3339–3352, 2012.

[32] M. V. Srinivasan, S. B. Laughlin, and A. Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond B Biol Sci.*, pages 427–59, 1982.

[33] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[34] Z. Wang and A. C. Bovik. Reduced and no–reference image quality assessment. *Signal Processing Magazine, IEEE*, 2011.

[35] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, 2003.

[36] M. H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *PAMI, IEEE Transactions on*, pages 34–58, 2002.

[37] P. Ye and D. Doermann. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7):3129 – 3138, July 2012.

[38] C. Zhang and Z. Zhang. A survey of recent advances in face detection, 2010.

[39] X. Zhu and P. Milanfar. A no-reference sharpness metric sensitive to blur and noise. In *QoMEX*, July 2009.