New Vistas in Image and Video Quality Assessment

Kalpana Seshadrinathan and Alan C. Bovik
The Laboratory for Image and Video Engineering (LIVE)
The University of Texas at Austin, Austin, TX-78712

ABSTRACT

In this Keynote Address paper, we review early work on Image and Video Quality Assessment against the backdrop of an interpretation of image perception as a visual communication problem. As a way of explaining our recent work on Video Quality Assessment, we first describe our recent successful advances on QA algorithms for still images, specifically, the Structural SIMilarity (SSIM) Index and the Visual Information Fidelity (VIF) Index. We then describe our efforts towards extending these Image Quality Assessment frameworks to the much more complex problem of Video Quality Assessment. We also discuss our current efforts towards the design and construction of a generic and publicly-available Video Quality Assessment database.

1. INTRODUCTION

The past several years have seen a resurgence in interest in *Perceptual Image Processing*, where algorithms for image processing are designed with human visual perception in mind. This is, of course, a very natural idea, but one that has met with limited success owing to our imperfect knowledge of the intended receiver, and indeed, of the transmitter. The receiver in this context, of course, is the remarkably complex human eye and visual cortex, while the transmitter we may take to be the environment, which casts images of extraordinary variability onto camera and retinal sensors.

The central idea behind perceptual image processing is to create, correct, or enhance images so that they have a visual appearance that is of highest *visual quality* to an average human observer. Of course, this statement is wrought with vagueness, since what does quality mean? Is it aesthetic appearance? Or is faithful rendering of the structure objects in a scene? Or is it maximization of scene information?

In any case, central to these pursuits is the question of *image and video quality assessment* (IQA and VQA), and whether algorithms can be developed that can successfully measure "quality" in an objective and perceptually meaningful manner. Certainly, applications exist where image quality is necessary for non-human interpretation, e.g., machine vision, but currently the end user in the overwhelming majority of image and video applications is a human observer. Moreover, following the notion that most of the great strides in machine vision research have derived from observations on biological vision, it is likely that quality assessment for the broad range of applications will benefit most greatly from studies that are relevant to human perception.

Thus, we are concerned with IQA and VQA algorithms that attempt to assess *perceptual degradations* in images and video signals. It is expected that success in this area will impact a wide variety of applications, since considerable gains in resource allocation should be achievable by using perceptual metrics for quality control. Indeed, judging by the authors' current interactions with industry, there is now an amazing interest in image and video quality. Of course, the tremendous proliferation in digital video products, ranging from wireline HDTV devices to portable video cell phones is driving this intense effort. Sophistications in image and video processing acquisition, compression, bandwidth enhancement for transmission over wired and wireless networks, the Internet, fast processing and display miniaturization have made possible the flood of current and promised video-based products. Such classic problems as de-noising, enhancement, restoration, error concealment and so on, are finding new applications in the mass-market consumer mainstream.

It is instructive to consider the following broad classification of the applications of quality assessment algorithms. In our view, IQA and VQA algorithms can be used:

- **To monitor video quality for real-time applications**. IQA and VQA algorithms can, in principle, be used to dynamically monitor and adjust the quality of video signals. For example, in video teleconferencing and Video

on Demand applications, video quality is affected by such factors such as errors, congestion and latency in the network, the number of participants in the multimedia stream, and so on. On-line quality monitoring has great potential to allow service providers to meet their Quality of Service (QoS) requirements by dynamically changing the resource allocation strategies.

- **To evaluation competing image and video processing algorithms**. A long-time bugaboo in the field of image processing has been the lack of any reasonable means for comparing algorithms. Indeed, for decades, the standard approach to assessing the relative merits of image processing algorithms has been to either "try them on Lena and see how they look," or to deploy a standard distance metric (relative to an idealized image) such as the Mean-Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR). The first of these approaches has the virtue of using the ultimate judge – the human observer, but suffers from a lack of statistical significance in the face of subjective variability, unless a massive and cumbersome human study is done encompassing adequate variety and numbers of images and observers, under controlled conditions, and with statistical validation of the results. Needless to say, such studies are very rare. The second approach has the virtues of implied automation and complete objectivity, but unfortunately, the objective measures used (MSE, PSNR, and so on) are acknowledged to have poor perceptual relevance, except, perhaps, at very low and very high error levels. Obviously, the emergence of truly successful IQA and VQA algorithms implies the possibility of changing this situation, and it is our hope that IQA/VQA algorithms will become standard impartial arbiters to measure and compare the efficacy of competing image and video processing algorithms.

- **To optimization the design of image and video processing algorithms.** A dual problem to testing the success of an image processing system using an IQA or VQA algorithm, is to design the system, or at least adjust its parameters, using a QA measure as the optimality criteria. Such systems could then be described as *Perceptually Optimal* if the IQA/VQA algorithms being used correlate sufficiently well with human judgment. Such an approach could revolutionize the design of image and video compression, filtering, acquisition, display (and so on) systems.

Regardless of how well we design IQA/VQA algorithms, controlled studies of the human subjective judgment of quality must remain the ultimate standard of performance for any image/video processing algorithm (except in machine vision, which is a much smaller market). Indeed, subjective judgment is the means by which IQA and VQA algorithms must be assessed. Subjective judgment involves psychophysical studies where image/video signals are viewed by human observers under controlled conditions. The subjects indicate a quality score on a numerical or qualitative scale. To account for human variability and to assert statistical confidence, multiple subjects are required to view each image/video, and a Mean Opinion Score (MOS) is computed. While subjective studies are the only completely reliable method of quality assessment, they are cumbersome, expensive, and complex [5]. Indeed, subjective QA is impractical for nearly every application – other than benchmarking automatic QA algorithms, for which it is an absolute necessity.

Objective image and video quality assessment algorithms are commonly categorized roughly into three types – namely full reference, reduced reference and no reference algorithms. We'll review these briefly, making some comments on their definitions.

Full Reference (FR) IQA/VQA algorithms make use of an ideal "reference" image that is assumed to be available for comparison. Naturally, this makes the problem easier! Indeed, nearly all of the work on IQA/VQA over the past few decades is FR, because of the relative simplicity of making quality judgments relative to a standard. FR algorithms are quite valuable, for all of the bulleted purposes listed above; yet, naturally, we'd like to do away with the need for reference data to achieve greater freedom of application, to develop insights into the meaning of visual quality, and simply because it has not yet been done.

Reduced Reference (RR) IQA/VQA algorithms do operate without the use of a pristine reference image or video, but they do make use of additional (side) information along with the distorted image or video signal. RR algorithms use typically use information or features *extracted* features from an original reference as supplemental comparison information, such as localized spatio-temporal activity information, detected edge locations, or embedded marker bits to estimate the distortion of the channel [1], [2], [32]. Other algorithms use knowledge that has been independently derived regarding the distortion process (such as foreknowledge of the nature of the distortion introduced by a compression

algorithm, e.g., blocking, blurring, or ringing) to assist in the quality assessment process [3]. Some authors refer to this type of RR algorithm as "blind," but in our view, presuming knowledge of the distortion process is a form of side information. RR techniques have attracted a lot of recent interest, since requiring the reference image/video may impose too much of a bandwidth limitation [4].

No Reference (NR) or *Blind* IQA/VQA algorithms attempt to assess image/video quality without access to any information than the distorted signal. In our view, algorithms that presume the distortion to be, for example, JPEG blocking, do not belong in this category. True NR IQA/VQA remains an exceedingly difficult, and indeed daunting proposition, and there is very little substantive work on this topic. Yet, it is certainly one of the "Holy Grails" of the image processing field, and the fact that human beings can perform the task almost instantaneously powerfully suggests that there is hope in this direction. Moreover, practitioners in industry are most interested in this problem, since evaluation of unknown video streams in a wireless cellular network undergoing a wide variety of compression, channel, and processing distortions affords little hope for using FR VQA algorithms in real-time, since the availability of a reference image is out of the question, while RR information is likely unreliable. Clearly, much remains to be learned regarding FR and RR QA and human perception of quality. No doubt, what is learned will eventually lead to feasible blind QA algorithms; while the problem at times seems nearly hopeless, and research on it remains moribund, we believe that the groundwork for eventual resuscitation of the area is being laid.

In the following Sections, we'll start by briefly reviewing early work on IQA and continue with some of our recent advances in IQA, specifically, the Structural SIMilarity (SSIM) Index and the Visual Information Fidelity (VIF) Index, which take two very different approaches to the problem. Both SSIM and VIF exhibit a remarkable level of performance relative to prior approaches to IQA. We will also describe recent successes we have had in extending both of these IQA frameworks to the much more complex problem of VQA. Our early work on VQA, while promising, still requires further algorithm development as well as more extensive testing resources. In later Section of this paper and the associated Talk, we discuss the our current efforts towards the design and construction of a VQA database, including details of the subjective study that needs to be conducted to complete the database. We plan to make this database available to the research community, as we have with our IQA database, which has become a standard in the research community. It is our hope that such a service will help us and others significantly advance the field of VQA.

2. BACKGROUND

It is useful in our discussion to broadly consider the process of a human observer viewing a displayed image using the philosophy of communication theory. Although we will refer to images in this context, the ideas involved extend directly to moving images, or video. In his framework, we may view the formation of images, which for the purpose of discussion we may regard as optical images of the natural or man-made environment under natural or standard man-made lighting, as *natural image transmission.* That is, the objects in the environment, along with the light sources, are the transmitters, and the lighted emitted from the objects (reflected or otherwise) is the *natural image signal.* We note in passing, as others have, that as with any engineered communication system, this signal has associated with it statistical properties that are a function of the source and the transmitter. These are commonly referred to as *natural scene statistics,* which have found utility in many image processing tasks.

These natural image signals radiate in all directions, and in so doing are modified by the ambient natural environment, which we might refer to as comprising the *natural image channel.* A small portion of this modified image signal is captured by image sensors, transduced into another electrical or optical form, then subsequently subjected to a series of intense processing steps which may include digitization, compression, modulation, channel coding, filtering, digital transmission, decompression, additional filtering, and subsequent display, the aggregate of all stages of which we may collectively refer to as the *synthetic image channel.* It is the overall natural-synthetic image channel, with all of its infinite complexities, that introduces distortion into the image signal. The great complexity of this overall channel, and the difficulty in generically modelling it, is one reason why blind IQA is such a difficult problem.

Finally, the displayed image signal is incident on the human eye, which along with the neurons along the visual pathway and the visual cortex, comprise what we may refer to as the *natural image receiver.* In the IQA literature, and indeed more broadly, the natural image receiver is referred to as the Human Vision System, or HVS.

Continuing with our analogy of imaging as a communication system in the classical sense, then we may also conclude that the more information that we have available regarding the nature of the transmitter, the channel, and the receiver, then the better job of image communication we will be able do, meaning the better quality images we will be able to efficiently deliver to the receiver, *provided that our models of transmitter, channel, and receiver are accurate,* and provided that we are able to effectively utilize this information in the design of the overall communication system. Key to this goal are the design of accurate and usable models of the natural image transmitters, the overall natural/synthetic image channel, and the natural image receiver.

Early *successful* modeling efforts focused on developing FR IQA and VQA algorithms using models of the natural image receiver - the HVS - to predict quality. The premise behind such HVS-based metrics is to simulate the visual pathway of the eye-brain system. In this framework, the error between a reference and test visual signal is computed in a *perceptual space*, in contrast to classical pure-math error metrics computed in the pixel domain, such as the MSE, the PSNR, and other similar distance metrics. As depicted in Fig. 1, HVS-based systems typically begin by preprocessing the signal to correct for non-linearities, since lightness perception is a non-linear function of luminance. A filterbank decomposes the reference and distorted ('test') image signals into multiple spatial frequency- and orientation-tuned channels in an attempt to model similar processing by the cortical neurons [6]. If the image signal is dynamic video, then a "temporal filtering block" as illustrated in Fig. 1 (using dashed lines) is typically deployed by HVS-based VQA algorithms, where the reference and test video sequences are also decomposed into temporal frequency channels. The luminance and contrast masking features of the HVS are then modeled to account for perceptual error visibility as a function of luminance and contrast. A space-varying threshold map is created for each channel describing local spatio-spectral error sensitivity, and is used to normalize the differences between reference and test images, resulting in what are referred to as *Just Noticeable Differences* (JND's). In the final stage, the JND values for all channels are pooled via a suitable metric such as a weighted MSE to generate a space-varying quality map.

This approach to quality assessment is intuitive and has met with considerable success. Indeed, we would agree that the idea of utilizing receiver models is the most natural approach to the problem of IQA/VQA. Popular IQA algorithms that followed the above paradigm include the pioneering work by Mannos and Sakrison [7], Lubin's laplacian-pyramid-based approach [8], Daly's Visible Differences Predictor [9], Teo and Heeger's steerable pyramid approach [10], and the Emmy Award-winning Sarnoff JNDMetrix metric [11]. Popular VQA algorithms following the HVS paradigm include the Perceptual Distortion Metric (PDM) [12], the Digital Video Quality (DVQ) metric [13], and the Sarnoff metric for video [11].

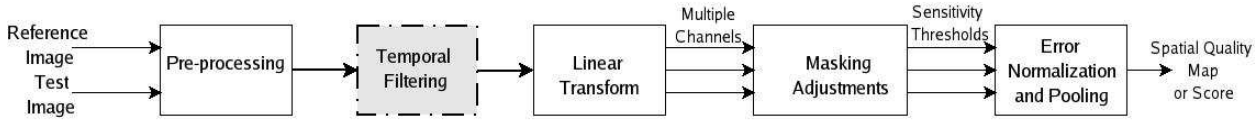


Figure 1. Block diagram of HVS-based metrics.

HVS-based metrics have several drawbacks that are well documented [16]. However, the main drawback, in our view, lies in the fact that our knowledge of the natural receiver, the HVS, while improving, still remains very limited. We have little idea, really, what the term *visual quality* really means. By contrast, we have a very good idea of what *visual fidelity* means, provided that we have an effective objective measure of fidelity. The hope is that HVS-based or hybrid image fidelity metrics will eventually be developed which correlate very highly with subjective IQA. Indeed, as our knowledge of human visual function increases, we believe this will very likely be the case.

While our incomplete knowledge of the natural receiver has apparently put limits on the performance of purely HVS-based IQA algorithms, the introduction of better-performing IQA algorithms has cast a strong light on these limits. One non-HVS-based algorithm that immediately attracted considerable attention is the Structural Similarity Index, or SIMM [16], which in an early form was called the University Quality Index, or UQI [17]. SSIM, which was one of the algorithms developed by our group, has found a great deal of popularity and visibility owing to its simplicity of definition, its ease of computation, its analytical tractability, and most of all, its excellent performance relative to human subjective scores (see Table 1). Interestingly, the idea behind UQI/SSIM arose not from systematic studies or modeling of the natural transmitter, channel, or receiver, but rather, from simple, intuitive ideas regarding how distortions in *image structure* might be measured. Indeed, the original UQI, or Wang-Bovik Index, arose in a back-and-forth Eureka

email exchange between the two inventors, both of whom were amazed by the excellent IQA performance of the algorithm.

Our group at UT-Austin had been studying IQA for some time prior to the development of UQI/SSIM, but with the unexpected success of UQI/SSIM, this work intensified and accelerated. In an effort to place IQA on a solid, and new theoretical footing, we chose to utilize recently-developed models of natural scene statistics to prototype the natural transmitter in our communication system analogy of IQA, which made possible the measurement of image fidelity in a natural information-theoretic setting. The resulting paradigm, known as the Visual Information Fidelity (VIF) Index [20], though more complex than SSIM in both construction and in computation, proved to be extremely efficient in terms of performance relative to subjective judgements (Table 1). VIF, like SSIM, had a prior simpler formulation which encapsulated the basic concepts of the new approach. The earlier algorithm, known as the Information Fidelity Criterion, or IFC [21], succeeded quite well but was improved upon in the later VIF formulation through the use of perceptual noise modeling and a form of divisive normalization [1], [21], [32].

Video Quality Assessment, or VQA, has followed a similar developmental trajectory as IQA. HVS-based metrics for video enjoyed considerable apparent success, until a VQEG study cast doubt the merits of the approach relative to simpler measures (such as the PSNR). All three video quality metrics mentioned above (PDM, DVQ, and Sarnoff) were proponents in a VQEG evaluation conducted as part of the Phase I FR-TV study in 2000 [14]. This study concluded that the performance of all proponents were, essentially, statistically equivalent to one another and to PSNR! The VQEG conducted another study in 2003, labeled Phase-II FR-TV study, to obtain finer discrimination between models than the Phase-I study [15]. Although the proponent models performed better in this study than in Phase-I, the Phase-II study emphasized a specific, and hence limited application domain, focusing on digitally encoded television. In this study, no single model emerged as a leading candidate. In our view, VQA remains an open area of inquiry where we expect that considerable strides might be made by exploiting paradigms, such as those motivating SSIM and VIF, that are complementary to those that only seek to model the natural receiver. New approaches to tackling the problem (we describe our own ideas in Section 4) will hopefully lead to viable alternatives that demonstrate generalizable VQA across a wide array of video distortions, with statistically superior performance relative to PSNR and other existing objective VQA algorithms.

3. IMAGE QUALITY ASSESSMENT

We briefly review the construction and performance of the SSIM and VIF Indices for still images as a way of introducing the concepts that will be extended into the video domain.

3.1 Structural Similarity Index

The UQI/SSIM Index tacitly assumes that the HVS has evolved to extract *structural information* from an image [16],[17]. Thus, the approach is, at least conceptually, a *dual approach* to HVS-based methods. The perceptual quality of a given image is predicted by quantifying the loss of structural image information, which is measured using simple sample statistics of image patches. Figure 2 illustrates the SSIM quality assessment system for images. Let $\mathbf{f} = \{f_i, i = 1, \ldots, N\}$ and $\mathbf{g} = \{g_i, i = 1, \ldots, N\}$ denote vectors from corresponding patches in the reference and test images **F** and **G**, respectively. From each patch, define weighted mean luminances and Root Mean Square (RMS) contrast. Also, define the covariance between the reference and test patches using:

$$\mu_{\mathbf{f}} = \sum_{i=1}^{N} w_i f_i , \qquad \sigma_{\mathbf{f}} = \sqrt{\sum_{i=1}^{N} w_i \left(f_i - \mu_{\mathbf{f}} \right)^2} , \quad \sigma_{\mathbf{fg}} = \sqrt{\sum_{i=1}^{N} w_i \left(f_i - \mu_{\mathbf{f}} \right)\left(g_i - \mu_{\mathbf{g}} \right)} ,$$

with similar definitions for $\mu_{\mathbf{g}}$ and $\sigma_{\mathbf{g}}$, and where the unit-sum weighting function w_i has a gaussian-like fall-off from the patch center. The SSIM Index between the corresponding patches is then

$$SSIM(\mathbf{f},\mathbf{g}) = l(\mathbf{f},\mathbf{g}) \cdot c(\mathbf{f},\mathbf{g}) \cdot s(\mathbf{f},\mathbf{g}) = \left(\frac{2\mu_{\mathbf{f}}\mu_{\mathbf{g}} + C_1}{\mu_{\mathbf{f}}^2 + \mu_{\mathbf{g}}^2 + C_1} \right) \cdot \left(\frac{2\sigma_{\mathbf{f}}\sigma_{\mathbf{g}} + C_2}{\sigma_{\mathbf{f}}^2 + \sigma_{\mathbf{g}}^2 + C_2} \right) \cdot \left(\frac{\sigma_{\mathbf{fg}} + C_3}{\sigma_{\mathbf{f}}\sigma_{\mathbf{g}} + C_3} \right)$$

where C_1, C_2, C_3 are small positive constants that stabilize each term. Thus, SSIM measures differences between patch luminances *l*, contrast *c*, and structure *s* expressed as simple, easily-computed image statistics. By computing *SSIM*(**f**, **g**) over the entire image, a quality map is obtained. A scalar quantifying the overall quality of the test image **G** is obtained by computing the overall mean value of *SSIM*(**f**, **g**), or by using some other pooling procedure [1], [31], [32]. The original UQI is the special case of SSIM where $C_1 = C_2 = C_3 = 0$.

Despite its simplicity, SSIM correlates extraordinarily well with perceptual image quality, and handily outperforms prior state-of-the-art HVS-based metrics such as the Sarnoff model, as demonstrated in extensive psychometric studies [1], [16], [22]. Table 1 shows the Spearman Rank Order Correlation Coefficient (SROCC) computed between the results of several metrics and the MOS images distorted by a wide variety of processes (the entire LIVE database) including JPEG and JPEG2000 compression, additive white gaussian noise, gaussian blur and fast-fading bit errors in a Rayleigh communication channel [16], [22]. The degree of improvement obtained by SSIM relative to the prior standard-bearer is nearly equal to the progress made over the previous thirty years of research! It should be noted that an improvement of 2-3%, while numerically small, is quite substantial from a perceptual standpoint. For example, the LIVE database contains distorted images with a wide range of degrees of each type of distortion, ranging from images in which the visual content is almost completely obscured, to images where the distortion is near the threshold of visibility. At these extremes, all of the algorithms correlate well with subjective judgment as might be expected (they'd better!). If the extreme distortions are left off, the disparity in performance between the algorithms becomes even more pronounced.

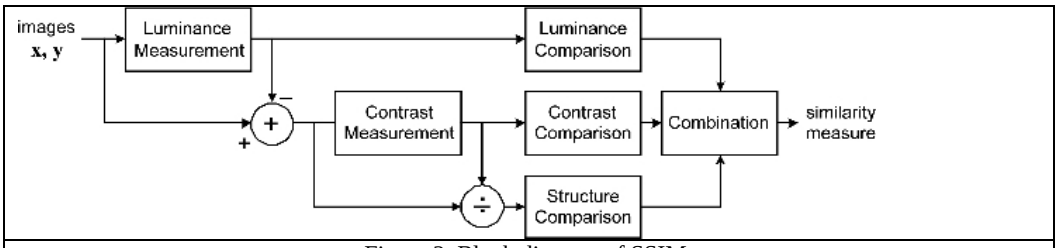


Figure 2. Block diagram of SSIM.

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	0.9263	0.8549	0.8779	0.7708	0.9854	0.7823	0.8907	0.8755
JND	0.9646	0.9608	0.9599	0.9150	0.9487	0.9389	0.9045	0.9291
DCTune	0.8335	0.7209	0.8702	0.8200	0.9324	0.6721	0.7675	0.8032
PQS	0.9372	0.9147	0.9387	0.8987	0.9535	0.9291	0.9388	0.9304
NQM	0.9465	0.9393	0.9360	0.8988	0.9854	0.8467	0.8171	0.9049
Fuzzy S7	0.9316	0.9000	0.9077	0.8012	0.9199	0.6056	0.9074	0.8291
BSDM (S4)	0.9130	0.9378	0.9128	0.9231	0.9327	0.9600	0.9372	0.9271
SSIM(MS)	0.9645	0.9648	0.9702	0.9454	0.9805	0.9519	0.9395	0.9527
IFC	0.9386	0.9534	0.9107	0.9005	0.9625	0.9637	0.9556	0.9459
VIF	0.9721	0.9719	0.9699	0.9439	0.9828	0.9706	0.9649	0.9584

Table 1: Performance of image quality metrics

3.2 Visual Information Fidelity Criterion:

The Visual Information Fidelity (VIF) Index views image quality assessment as an *information fidelity* problem. The philosophy of the approach is based on the hypothesis that visual quality is related to the amount of information that the HVS can extract from an image. In this sense, VIF is also a dual approach to IQA relative to HVS-based methods.

Figure 3 summarizes the VIF approach. Reference images are assumed to be the output of a natural image source represented using a powerful, yet simple natural scene statistic (NSS) model known as the Gaussian Scale Mixture (GSM) model [19]. In this model, the wavelet coefficients of the natural source **f** have a GSM distribution: the elements **f** of a patch of spatially adjacent locations, scales and orientations are distributed as a zero-mean Gaussian random vector, conditioned on a multiplier field: $\mathbf{f} \sim z\mathbf{u}$, where *z* is a scalar multiplier field, and **u** is a zero-mean Gaussian random vector with covariance matrix $\mathbf{C_u}$. The multiplier field is estimated from the reference image, while the test

image **g** is assumed to be the output of a distortion channel through which the reference image passes. A blur plus additive noise distortion model in the wavelet domain is used as the channel model.

If **g** are the corresponding coefficients from the test image, then $\mathbf{g} = b\mathbf{f} + \mathbf{n}$ where b is a scalar gain field that models modification of signal energy due to compression, blur, additive noise, contrast enhancement and/or other distortion, and **n** is zero-mean AWGN. Further, the HVS is modeled as a zero-mean AWGN communication channel **v**, since neural noise and other factors limit the information it can extract from an image. Thus, the "perceived" reference and test images are (respectively) $\mathbf{e} = \mathbf{f} + \mathbf{v}$ and $\mathbf{d} = \mathbf{g} + \mathbf{v}$.

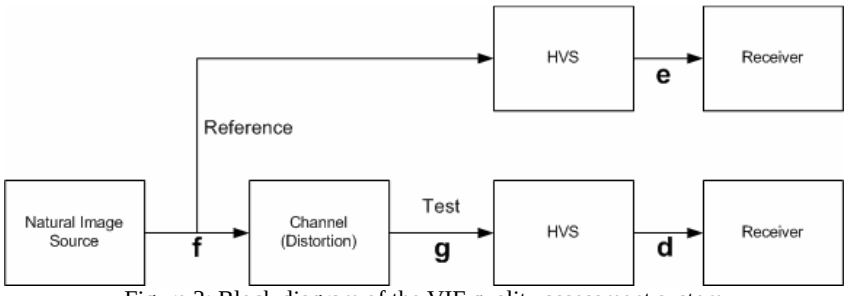


Figure 3: Block diagram of the VIF quality assessment system

The mutual information I between **f** and **d** (conditioned on z) measures the information that the HVS can extract from the test image **g** and likewise for the reference image **f**. The ratio of these information measures defines the *local* VIF Index of the distorted image:

$$\mathrm{VIF}(\mathbf{f}, \mathbf{g}) = \frac{\mathrm{I}(\mathbf{f}; \mathbf{d} \mid z)}{\mathrm{I}(\mathbf{f}; \mathbf{e} \mid z)}$$

As with SSIM, a global VIF quality metric can be obtained by summing both numerator and denominator over all patches and all wavelet subbands, then forming the ratio. A detailed discussion of VIF, including parameter and wavelet basis selection can be found in [20]. The precursor to VIF, known as the *Information Fidelity Criterion*, is described in [21]. This metric is similar to VIF, but does not include normalization with respect to the information content (and does not perform as well).

The performance of VIF has been tested extensively across widely varying distortion types and found to exhibit superior performance relative to all known algorithms (including SSIM) as indicated, for example, using SROCC relative to MOS as a performance metric (Table 1). An extensive description of this study, which included over 25,000 value judgments on nearly 800 distorted images, and which deployed a wide variety of statistical measures of algorithm performance, is available in [22].

4. VIDEO QUALITY ASSSESSMENT

Most of the video quality metrics proposed in the literature have been simple extensions of still image quality metrics, or IQA algorithms. Thus, most of the leading VQA algorithms are based on models of the natural receiver, the HVS, as discussed in Section 2. In fact, seven of the ten proponent models evaluated by the VQEG in its Phase I testing used models of the HVS in their algorithms [14]. More recently, there has been a shift in trend toward so-called *top-down approaches* to quality assessment, which we can equate to modeling of either the natural image transmitter, and/or the natural/synthetic image channel, using our prior parlance. As we mentioned before, part of the reason for this shift in paradigm has been the realization that we are still limited in our understanding of the complexity of the HVS, which means that our current HVS models contain inaccuracies that may degrade the performance of HVS-based VQA algorithms.

A more important reason, perhaps, is the fact that HVS-based models typically model *threshold psychophysics*, *viz.*, the sensitivity of the HVS to different features such as luminance, contrast and contrast masking phenomena are measured at the threshold of perception [23]. However, quality assessment usually deals with supra-threshold perception, where

artifacts in the video sequences are visible and algorithms attempt to quantify the annoyance levels of these distortions. Thus, in recent years, there has been an increased interest in models that describe the distortions in the video sequence that the human eye is sensitive to and that equate with loss of quality; for example, blurring, blocking artifacts, fidelity of edge and texture information in the signal, color information, contrast and luminance of registered patches in the spatial and frequency domain, and so on. These approaches, in our framework, would be described as using implicit models of the natural image transmitter, and explicit models of the natural/synthetic image channel. Indeed, five of the six proponent models tested by the VQEG in its Phase II testing utilized feature vectors that contained information such as those just described in predicting quality [15]. These feature vectors are generally combined either using linear weighting measures or non-linear learning mechanisms to compute a quality index for the entire sequence. Although these developments are promising, in particular for specific video industries, we note that the reliance on models of particular distortions is likely to limit the general application of these methods.

Another problem, in our view, is that the top-down algorithms for VQA proposed in the literature have incorporated features for measuring *spatial distortions* in video signals, yet very little effort has been spent on measuring *temporal distortions* or motion artifacts. Several of the algorithms mentioned above utilize rudimentary temporal information by differencing adjacent frames or by processing the video using other temporal filters before feature computation. However, to our knowledge, no algorithms in the literature attempt to compute *motion information* in video signals to predict quality. Yet, The HVS is quite sensitive to motion and can accurately judge the velocity and direction of moving objects. These skills are essential to survival and play a huge role in human perception of moving image sequences. Considerable resources in the HVS are devoted to motion perception. Most of the neurons in the striate cortex respond best to a stimulus moving in a particular direction and play a role in the perception of movement. Motion perception is largely executed in the medial temporal (MT) area of extra-striate cortex, where 90% of the neurons are directionally sensitive. Motion perception is not modeled well in current HVS-based design paradigm either. All of the VQA metrics mentioned above use either one or two temporal channels, and model the temporal tuning of the neurons in area V1 of visual cortex only – despite the important role of the neurons in area MT of the extra-striate cortex in motion perception.

We believe that the performance of VQA techniques can be improved by the introduction of meaningful models that describe motion in video sequences, as well as model spatio-temporal distortions in the video stream. To date, there has been very little work done in these directions. Next we will describe video quality metrics that we have very recently developed that are based on the structural similarity and visual information fidelity concepts. The novelty in this new work lies in incorporating one of the chief factors that affect human perception of moving image sequences – namely, the motion of objects in the scene that is viewed.

4.1 Basic framework for motion modeling

The motion of objects in 3D scenes takes an elegant and simple form in the frequency domain that facilitates analysis. We consider the apparent motion of image intensities, namely the optical flow, and not the true three-dimensional velocity of motion. We assume that short segments of video consist of local image patches undergoing translation, which is a reasonable approximation as long as there are no scene changes. This model is used *locally* to describe video sequences, since translation is a linear approximation to more complex types of motion. Let $i(x,y)$ denotes an image and $\tilde{I}(u,v)$ its Fourier transform. Assuming that this image undergoes translation with flow vector $\vec{\lambda}=(\lambda_x, \lambda_y)$, the Fourier transform of the resulting video sequence, denoted $F(u,v,w)$, is given by [27]:

$$\tilde{F}(u,v,w) = \tilde{I}(u,v)\,\delta(\lambda_x u + \lambda_y v + w) \tag{1}$$

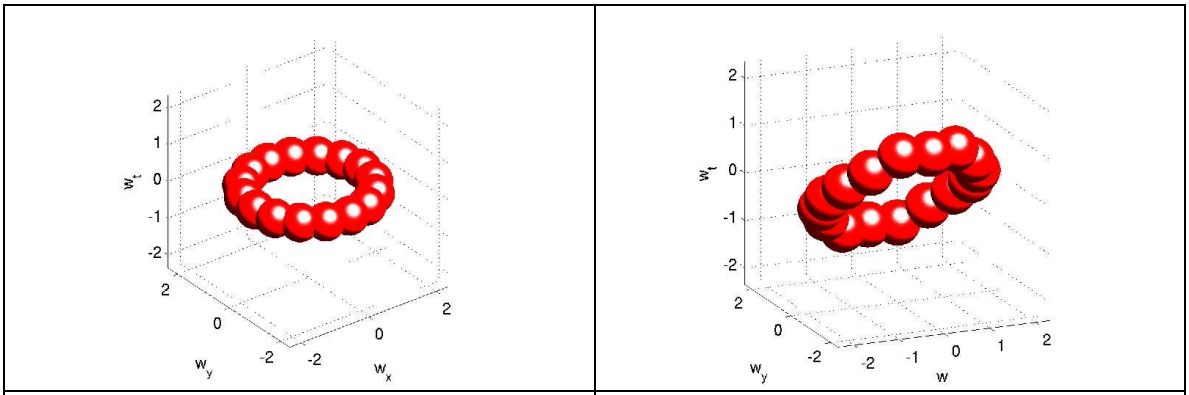
where $\delta(x)$ denotes the Dirac delta function. Thus, the spectrum of a translating video signal lies entirely along a plane in the frequency domain whose orientation is defined by the flow vector. Additionally, the magnitudes of the spatial frequencies do not change, but are simply sheared in the frequency domain.

Eq. (1) provides an explicit characterization of the motion of a video sequence in the frequency domain. Frequency domain approaches are well suited to the study of human perception of video signals owing to the presence of bandpass visual channels in the HVS [5]. Hence, in the VQA systems that we describe, the video sequence is filtered using a family of band-pass spatio-temporal filters and quality assessment is performed on the resulting bandpass channels in

the spatio-temporal frequency domain. Although the development is generic and applies to any filter family, we use Gabor filters in our implementation to achieve improved spatio-spectral localization [28].

We used the optical flow estimation algorithm described in [29] to compute the flow field for the reference video, with small modifications that are described in [25]. The optical flow computation on the reference sequence provides us with an estimate of the local orientation of the plane containing the frequency spectrum of the video sequence. We propose to use this information to achieve motion compensated quality computation in the following way. We first identify the Gabor filters that overlap significantly with this plane, which is accomplished by requiring that the plane lie within one standard deviation of the Gabor filter in the frequency domain. This is illustrated in Fig. 4 for two hypothetical video sequences, where one is a static sequence that contains no motion and the other is a moving sequence. Note the close relation between using this rule for filter selection and motion compensated filtering of the reference sequence. If the video signal exactly satisfies our assumption of translational motion, then exact motion compensated filtering of the video sequence would require using a filter whose support lies entirely along the spectral plane of the reference video signal. This can be achieved using non-separable Gabor filters, whose major axis is oriented along the spectrum of the video and whose spread along the minor axis is very small or negligible. We, however, use spherically symmetric Gabor filters that are separable. The filtering that we achieve is therefore not exactly motion compensated, but can be considered an approximation to it, provided that the spread of the Gaussian window in the frequency domain is not too large.

Once these filters have been identified, we compute quality indices only between the outputs of the selected subset of filters. This helps us to model both spatial as well as temporal distortions in the test video sequence. Distortions in the video that are *purely spatial*, meaning intra-frame distortions, will result in changes in the frequency components along the plane, which will be captured by the Gabor filter outputs. Examples of such spatial distortions include blurring, blocking and ringing caused by compression, errors during acquisition, transmission through communication channels, and so on. Distortions in the video that are *purely temporal*, meaning inter-frame distortions, will result in a change in the axis along which the plane intersects the Gabor filter. Examples of temporal distortions include motion compensation mismatch and mosquito noise due to compression, ghosting and temporal aliasing during acquisition, transmission through communication channels etc. We describe a specific instance of this generic framework for VQA in the following.



(a) (b)

Fig. 4: Illustration of a set of filters chosen for the case of (a) A static sequence (video consists of same frame repeated throughout the sequence) (b) A sequence undergoing translation

4.2 Video Structural SIMilarity index (V-SSIM)

Soon after the development of SSIM, a very simple extension of SSIM was proposed for VQA, wherein a simple frame-by-frame SSIM implementation was described [24]. Although this algorithm proved to be competitive with the VQEG proponents from [14], we have been able to demonstrate that the SSIM paradigm can yield much better performance by

introducing distortion measurement that incorporates motion information [25]. The development of this approach, which we term Video SSIM, or simply the V-SSIM Index closely follows the development of the Complex Wavelet SSIM (CW-SSIM) Index proposed in [26]. CW-SSIM is a simple extension of SSIM, where structural similarity is measured in the complex wavelet domain, thereby achieving high performance and a degree of translation-invariance, which is quite useful if errors in registration occur between the reference and test video sequences.

V-SSIM is defined as follows. Let $\vec{f} = \{f_i, i = 1, 2, \ldots, N\}$ and $\vec{g} = \{g_i, i = 1, 2, \ldots, N\}$ denote sets of coefficients from the reference and distorted video sequences, respectively, at corresponding spatio-temporal locations from one sub-band of the Gabor filter family. Then, the V-SSIM Index between these coefficients is given by:

$$\text{V-SSIM}(\vec{f}, \vec{g}) = \frac{\sum_{i=1}^{N} |f_k \, g_k| + K}{\sum_{i=1}^{N} |f_k|^2 + |g_k|^2 + K} \tag{2}$$

where K is a small positive constant added to preserve numerical stability. Note that we use only the magnitudes of the Gabor filter outputs to compute the V-SSIM Index, by contrast with the CW-SSIM Index [26]. Since the CW-SSIM Index was designed using the complex wavelet response in order to yield a translation insensitive measure, the phase of the complex wavelet response corresponds to small translations in the image. However, for application to VQA, the phase of the Gabor outputs represents *motion information*, and the optical flow estimation algorithm in [29] computes flow using this phase information. Thus, once motion compensation is accomplished, then the V-SSIM Index is only computed between the magnitudes of the *selected* filter outputs, using the selection criterion described in Section 4.1.

We tested our proposed V-SSIM index on the VQEG database [14] and the results we report are from [25]. We are not particularly satisfied with the VQEG database in terms of the types of video sequences and distortions that are represented, but it is the best available *current* VQA database. Nevertheless, as described below, we have plans to create, in the future, a VQA database to complement the existing popular LIVE IQA Database. The VQEG database contains 20 reference video sequences, test sequences obtained by distorting each of these reference videos with 16 different distortion operations and subjective scores for all test sequences. The current implementation of our optical flow estimation uses filters at just one scale. Therefore, we had to exclude 4 of the reference sequences in the database that contained fast moving sequences, where the flow estimation algorithm failed due to temporal aliasing [29]. All of the VQEG test sequences are interlaced and for simplicity, and to avoid the degrading effects of applying de-interlacing, our algorithm operates only on the odd fields of the interlaced sequences. The results of our simulations are summarized in Table 2, which shows the Spearman Rank Order Correlation Coefficient (SROCC) between the subjective and objective scores for several different VQA algorithms. SROCC is one of the metrics specified by the VQEG that tests the prediction monotonicity of a VQA system. As can be seen, and as expected, the PSNR does not correlate well with subjective scores. Proponent P8 is the best performing metric amongst the 10 different proponent models tested by the VQEG in terms of SROCC [14]. We also compared our results against the better performing version of the two metrics proposed in [24]. As can be seen, the V-SSIM Index is attractively competitive with the leading VQA algorithms.

Prediction Model	SROCC
PSNR	0.786
Proponent P8 (Swisscom) in [14]	0.803
Frame-by-frame SSIM in [24]	0.812
V-SSIM using motion [25]	0.835

Table 2

4.3 Video Information Fidelity

We recently proposed a model that describes the statistics of natural video sequences, towards the development of an information theoretic quality metric for video signals [28]. Translational motion of local image patches was combined

with the GSM model for natural images in the frequency domain to describe the statistics of sub-band filtered coefficients of video signals. Assume the video signal $\boldsymbol{f}$ is filtered with a family of spatio-spectrally localized 3-D subband filters $g_i(x, y, t) \leftrightarrow \tilde{G}_i(u, v, w)$, resulting in wavelet coefficients $c_i(x, y, t)$. A modified version of the GSM model for natural images is used to capture the behavior of the image undergoing translation, and an associated mixing density parameter z is used as described in [28]. When conditioned on the estimated value $\hat{z}$ of the mixing density z, the coefficients $c(x,y,t)$ are zero-mean Gaussian random variables with variances:

$$\sigma_c^2 = (2\pi)^{-6} \int \hat{z}^2 \left| \tilde{G}\left(u, v, -\lambda_x u - \lambda_y v\right) \right|^2 du\, dv$$

where $\tilde{G}\left(u, v, -\lambda_x u - \lambda_y v\right)$ is a 2-D slice of the filter along the plane containing the spectrum of the translating video signal. From this it is apparent that large-magnitude coefficients will appear where the energy of the variance field is large, and where the oriented plane significantly intersects the filter passbands.

Our work on developing information-theoretic algorithms for VQA based on our prior work on IQA is still ongoing. Recently, we have developed a video analog of the IFC Index (VIF without perceptual noise modeling and divisive normalization), which appears to have performance that is favorably competitive with all prior VQA algorithms [30]. However, our perceptual testing of this algorithm remains incomplete as of this writing. Further, we expect to soon incorporate modifications, following the VIF Index, which will likely further improve our information-theoretic approach to VQA.

5. TOWARDS A VIDEO QUALITY ASSESSMENT DATABASE

The most important tool assisting our successful advancement of the field of Image Quality Assessment is the LIVE Quality Assessment Database, which has become a *de facto* standard in the global image processing community. The database contains 779 images - 29 reference images distorted by a diversity of processes such as JPEG/JPEG 2000 compression, blur, AWGN and wireless channel bit errors [22]. Each image was evaluated by (on average) 23 human observers to determine Mean Opinion Scores (MOS). The recent Release 2 of the LIVE Database includes Differential MOS (DMOS) values as well, which is regarded as more sensitive than MOS. Over 200 institutions have downloaded the LIVE database for research purposes – despite the data volume of >1GB - and it has already been cited in over 20 technical articles, although it was first was released to the research community less than two years ago. It is safe to claim that the testing of new IQA algorithms by researchers around the world is conducted on the LIVE database.

To enable the performance evaluation of VQA algorithms over a suitably diverse dataset, we plan to expand the LIVE database by developing a VQA database of generic power, containing videos affected by a broad variety of important and general distortions. We also plan to provide subjective scores (MOS, DMOS) for all of the distorted videos. There is a great need for a standardized VQA database that is freely available to the research community, and which go beyond the needs of specific video industries. Conducting subjective studies on many video sequences is a cumbersome and daunting task, but we believe that such an effort will prove to be a great service to the community, which encourages us to undertake this task.

Currently, VQA algorithms can be tested against other algorithms on the VQEG Phase-I FR-TV database. However, the VQEG database has significant limitations. First, 4 of the 5 current VQEG projects involve evaluation of VQA metrics for television systems. As a result, the reference and distorted videos in the database are interlaced, leading to visual artifacts in the *reference* as well as distorted video sequences. Algorithms such as those mentioned in Sections 2 and 4 typically involve multiple processing steps which require adjustment to handle interlaced signals. De-interlacing the sequence prior to processing is not suitable in a VQA framework, since de-interlacing introduces artifacts in the video. Additionally, the VQEG database consists *only* of compression-related artifacts produced by H.263 and MPEG codecs. This set of distortions is quite limited for the broad spectrum of distortions that generic quality assessment algorithms are targeted towards.

Thus, we plan to develop a database of progressive scanned videos suitable for a broad range of multimedia applications, such as video teleconferencing, Video-on-Demand, Internet streaming video, mobile multimedia, and so

on. We will incorporate a wide range of distortions, including *MPEG-2 compression*, which exhibits compression artifacts such as blur, ringing, motion compensation errors, mosquito noise etc.; *H.263 compression*, representative of low-bit rate video compression to be widely deployed in mobile applications; simulated spatial distortions including *additive noise* and *blur* and temporal distortions such as *jerkiness* and *smearing*; and *channel errors* including delay-induced *jitter*, visual artifacts from *error concealment* of lost packets, *bit errors* and *burst errors* due to fading and multi-path reflections, and so on.

A major difficulty that we (and others) have encountered is the acquisition of high-quality, progressive scanned, copyright free source videos, so that we can share the data as well as the result of our subjective studies freely with the research community. Towards this end we have, thus far, obtained 8 videos from a High Definition database provided by the University of Munich and several videos from the mobile video unit at Texas Instruments, Inc. These sequences contain horizontal, vertical, panning, zoom and rotary motions and both fast moving and slow moving scenes. Our eventual goal is about 20 videos.

Finally, we will undertake an extensive psychometric study in consultation with noted visual psychologists and our frequent collaborators L. Cormack and W. Geisler at UT-Austin's Center for Perceptual Systems (CPS) at UT-Austin. LIVE/CPS possesses ample resources for calibrated video display. We intend to conduct a Single Stimulus Continuous Quality Evaluation (SSCQE) procedure to compute scores for all sequences in the data sets. Such a study is well suited to applications such as video monitoring and quality control that are rapidly gaining popularity. Additionally, the SSCQE procedure allows for the subject to provide a *time-dependent* index of quality, as opposed to a single index that determines the quality of the entire signal. We envision that the resulting database would prove more challenging for VQA algorithms than the current VQEG database, and would enable more rigorous performance evaluation of quality assessment systems.

REFERENCES

1. Z. Wang and A.C. Bovik, *Modern Image Quality Assessment*. New York: Morgan and Claypool Publishing Company, 2006.
2. V.-M. Liu, J.-Y. Lin and K.-G. Wang, "Objective image quality measure for block-based DCT coding," *IEEE Transactions on Consumer Electronics*, vol. 43, pp. 511-516, June 1997.
3. S. Liu and A.C. Bovik, "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, pp. 1139-1149, Dec. 2002.
4. M. Masry, S.S. Hemami and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, pp. 260–273, 2006.
5. Winkler, *Digital Video Quality*. New York, Wiley, 2005.
6. A.C. Bovik, M. Clark and W.S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-12, pp. 55-73, Jan 1990.
7. J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974.
8. J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, A.B. Watson (Ed.), MIT Press, p. 163, 1993.
9. S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A.B. Watson (Ed.), MIT Press, p. 179-206, 1993.
10. P.C. Teo and D.J. Heeger, "Perceptual image distortion," *IEEE International Conference on Image Processing*, 1994.
11. Sarnoff Corporation, JNDMetrix Technology: http://www.sarnoff.com/products services/video vision/ jndmetrix/downloads.asp
12. S. Winkler, "Perceptual distortion metric for digital color video," in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 3644, no. 1, San Jose, CA, USA: SPIE, pp. 175–184, May 1999.
13. A.B. Watson, J. Hu, and J.F. McGowan III, "Digital video quality metric based on human vision," *J. Electron. Imaging*, vol. 10, no. 1, pp. 20–29, Jan. 2001.
14. Final VQEG report on the validation of objective quality metrics for video quality assessment: http://www.its.bldrdoc.gov/vqeg/projects/frtv phaseI/index.php

15. Final VQEG report on the validation of objective models of video quality assessment: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/downloads/VQEGII_Final_Report.pdf
16. Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.
17. Z. Wang and A.C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.
18. H.R. Sheikh, Z. Wang, L.K. Cormack, and A.C. Bovik. The live image quality assessment database: http://live.ece.utexas.edu/research/quality/
19. M.J. Wainwright and E.P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, S.A. Solla, T. Leen, and S.-R. Muller, Eds., vol. 12, 1999, pp. 855–861.
20. H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing* vol. 15, no. 2, pp. 430–444, 2006.
21. H.R. Sheikh, A.C. Bovik and G. DeVeciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117-2128, December 2005.
22. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.
23. T.N. Pappas, R.J. Safranek and J. Chen, "Perceptual criteria for image quality evaluation," Chapter 8.2 in *The Handbook of Image and Video Processing*, Second Edition, A.C. Bovik, (Ed.), New York: Elsevier Academic Press, pp. 939-960, 2005.
24. Z. Wang, L. Lu, and A.C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, pp. 121–132, Feb. 2004.
25. K. Seshadrinathan and A.C. Bovik, "A structural similarity metric for video based on motion models", To appear in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing* , Honolulu, Hawaii, May 2007.
26. Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, Pennsylvania, Mar. 2005.
27. A.B. Watson and A.J. Ahumada, "Model of human visual-motion sensing," *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 322–342, 1985.
28. K. Seshadrinathan and A.C. Bovik, "Statistical video models and their application to quality assessment," *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, January 23-25, 2006.
29. D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information", *International Journal of Computer Vision*, vol. 5, no. 1, pp. 44-104, Aug. 1990.
30. K. Seshadrinathan and A.C. Bovik, "An information theoretic video quality metric based on motion models," To appear in *Proc. Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, Jan. 2007.
31. Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," *IEEE Int'l Conf. on Image Processing*, Atlanta, GA, Sept. 2006.
32. A.C. Bovik, *The Handbook of Image and Video Processing*. Second Edition, Elsevier, 2005.