IMAGE AND VIDEO QUALITY ASSESSMENT

Sumohana Channappayya, Kalpana Seshadrinathan and Alan C. Bovik

1. Introduction

The reader will no doubt have observed that digital images and videos are playing an ever-increasingly pervasive part in our daily lives. They are making their way into our living rooms, laptops, hand-held devices, and cell phones. High resolution High Definition (HD) digital video broadcasts, as well as lower resolution streaming video over wireless networks are here to stay. The popularity of video-centric websites such as Youtube and Facebook are good examples of how this communication medium has impacted our lives. Indeed, acronyms like JPEG, MPEG, and H.264, once the parlance of engineers only, have become a part of our common vocabulary.

Given the phenomenal rate at which image and video content is being generated and distributed, a critical task is the evaluation of the perceptual quality of the content. For example, video content providers need to evaluate encoding parameters, while network service provider need perceptual quality scores to decide load balancing. Subjectively evaluating the quality of the content is an extremely difficult task due to the time and cost involved. Indeed, the only reliable subjective test involves using large numbers of human test subjects, under controlled psychometric experimental conditions, to evaluate the images and/or videos, resulting in statistically meaningful Mean Opinion Scores (MOS). This approach is of course impractical in most situations. The ideal substitutes for human subjectivity are *objective quality assessment algorithm*s whose scores have been shown to correlate highly with human subjectivity. Perfect correlation is, of course, impossible, since human subjects vary in their judgment too much. There are many challenges in the design of such algorithms for image and video quality assessment. In this article we discuss the challenges involved and present some state-of-the-art image and video quality assessment algorithms.

Generally, image and video quality assessment algorithms are classified into three groups – full-reference (FR), reduced-reference (RR), and no-reference (NR) algorithms. As their names suggest, the groups correspond to the amount of information available about the original, presumed pristine reference signal. The design of true no-reference algorithms is extremely challenging and little progress has been made. Reduced reference algorithms are somewhat easier and are interesting, but are generally specific to an application. In this article, we limit our discussion to full-reference image quality assessment (IQA) and video quality assessment (VQA) algorithms, where much progress has been made.

2. Image Quality Assessment

The primary goal is to produce automatic image and video rating that correlate well with MOS. A natural approach is to try to mimic the human visual system (HVS), but the HVS itself is still poorly understood. Several of these types of FR algorithms have been proposed, notably, the just noticeable difference (JND) metric. Another approach is to treat IQA as evaluation of an image communication system. This approach expresses the test signal as the reference signal distorted by an imperfect communication channel, and uses properties of both the source and receiver in its design. The communication system model has resulted in

the emergence of two popular IQA algorithms called SSIM and VIF. First however, we discuss a very popular - but flawed - measure of quality – the mean squared error (MSE).

2 (a) Mean Squared Error (MSE)

The MSE and the related peak signal to noise ratio (PSNR) are popularly used to assess image quality. Given two vectors $\mathbf{x} = \{x_i | i = 1, \ldots, N\}$ and $\mathbf{y} = \{y_i | i = 1, \ldots, N\}$, then

$$\mathrm{MSE}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2,$$

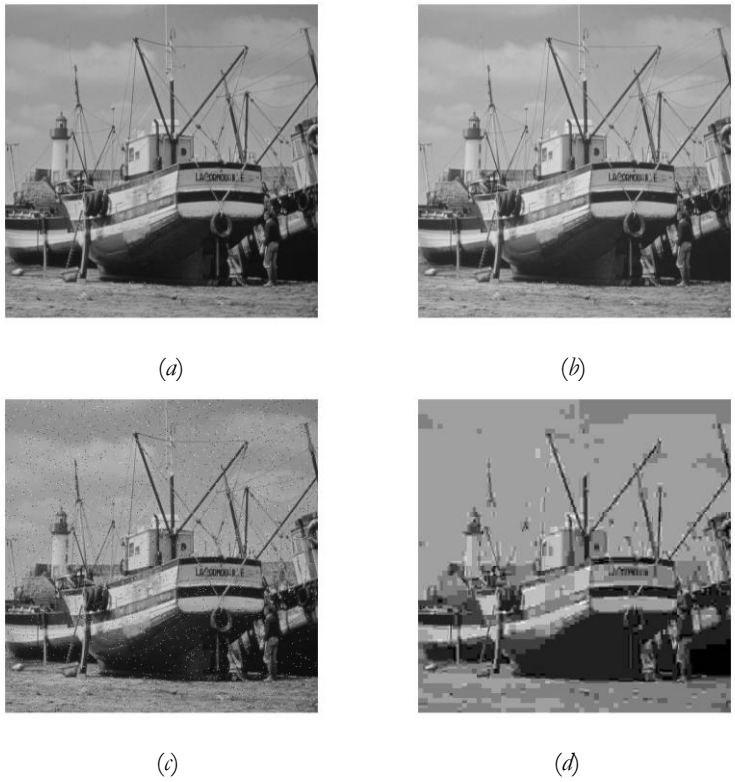
while

$$\mathrm{PSNR}(\mathbf{x}, \mathbf{y}) = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}(\mathbf{x}, \mathbf{y})} \right),$$

where L is the image dynamic range (typically [0, 255]).

Of course, the MSE is easy to compute and implement in software and hardware, and has not suffered from any competition until recently. Moreover, it is easy to use in analysis and often gives closed form solutions to optimization problems.

However, the MSE is a very poor measure of image quality! Specifically, it correlates very poorly with MOS. A simple illustration is shown in Fig. 1. A human would likely rate the distorted images in order (best) Fig. 1(b), 1(c), 1(d) (worst) relative to the reference. However, all three images have identical MSE of 235 relative to the reference!



(*a*) (*b*)

(*c*) (*d*)

Fig. 1: Example of inadequacy of MSE for measuring image quality. (a) Reference "Boats" (b) Mean shifted. (c) Distorted with salt and pepper noise. (d) JPEG compressed. All three (*b*)-(*d*) have the same MSE = 235!

However, the SSIM Index values for the three images are 0.98, 0.73, and 0.68, respectively. Moreover, the VIF Index scores are 0.99, 0.44 and 0.14, respectively.

2 (b) Structural SIMilarity (SSIM) Index

The SSIM index is a recent and very popular IQA algorithm that uses properties of source and the receiver in its design. The idea behind SSIM is that natural images are highly structured, and that the eye is sensitive to *structural distortion.* The SSIM index expresses quality by comparing local correlations in luminance, contrast, and structure between reference and distorted images. Given signal vectors **x** and **y**, the SSIM index is

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \mathrm{l}(\mathbf{x},\mathbf{y})\cdot \mathrm{c}(\mathbf{x},\mathbf{y})\cdot \mathrm{s}(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2+\mu_y^2+C_1}\cdot\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2+\sigma_y^2+C_2}\cdot\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3},$$

where l(**x**, **y**), c(**x**, **y**) and s(**x**, **y**) compare local image luminance, contrast, and structural correlation respectively. Also μ_x, μ_y are sample means, σ_x^2, σ_y^2 are sample variances, and σ_{xy} is the sample cross-covariance between **x** and **y**. The constants C_1, C_2, C_3 stabilize SSIM when the means and variances become small. A simplified form of SSIM is popularly used ($C_3 = C_2/2$):

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2+\mu_y^2+C_1}\cdot\frac{2\sigma_{xy} + C_2}{\sigma_x^2+\sigma_y^2+C_2}$$

In IQA image quality assessment, local patches from reference and distorted image constitute **x** and **y**. The average SSIM value over the image (*Mean SSIM* or *MSSIM*) gives the final quality measure. The MSSIM scores are also given in Fig. 1 – clearly more in accordance with perception! Of course, high correlation with MOS is what counts – which has been shown conclusively: see the paper "Image Quality Assessment: From Error Visibility to Structural Similarity" (*IEEE Trans. on Image Processing,* vol. 13, no. 4, Apr. 2004).

2 (c) Visual Information Fidelity (VIF) Index

The VIF Index is another recent IQA algorithm that consistently outperforms the MSE and the JND metric. It uses a fundamentally different design philosophy, treating IQA problem as an information fidelity problem, where the image is the information source, the distortions modeled by a channel, and the HVS the receiver. The source is modeled as the output of a stochastic process, and the channel as a combination of blur plus noise. The receiver is assumed noisy modeled by additive white Gaussian noise.

Image quality is then defined as the mutual information between the source and the distorted observation. Thus, image quality is a function of the distortion channel characteristics and mutual information is equal to the channel capacity.

We describe the VIF index in detail. First and foremost, the VIF index operates in a multi-resolution transform domain. The motivation is two-fold: (a) the HVS represents images at multiple resolution levels, and (b) good statistical models for multi-resolution image transform coefficients exist. Specifically, a Gaussian scale mixture (GSM) model is used. We refer to the transform coefficients as wavelet coefficients in the sequel.

Source characteristics: A length-M vector **c** composed of neighboring wavelet coefficients of the reference image is modeled as a GSM using the relation $\mathbf{c} = z\mathbf{u}$, where z is a scalar random variable ($z \geq 0$) and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{C_u})$ is a Gaussian random vector. z, **u** are independent. Thus the *conditional* distribution $f_c(\mathbf{c} \mid z)$ is Gaussian, which greatly simplifies the analysis. Intuitively,

z is the local standard deviation of wavelet coefficients scaling **u**. Since the VIF index is a FR algorithm, it uses the reference image to estimate z and **u**.

Channel characteristics: The VIF Index approximates distortions by a combination of blur plus noise. If **c** is a vector of wavelet coefficient from a given location in the reference image, and **d** is the corresponding vector from the test, then
d = g**c** + **v,**
where g represents blur and **v** is additive white Gaussian noise, **v** ~ $N(\mathbf{0}, \sigma_v^2\mathbf{I})$. In VIF, g and σ_v^2 are estimated from **c**.

Receiver characteristics: The receiver model is very simple: aside from the wavelet decomposition, the HVS is modeled only by AWGN neural noise: **n ~** $N(\mathbf{0}, \sigma_n^2\mathbf{I})$.

Putting these models together, the reference image perceived by the eye is
e = **c** + **n,**
and the test image is
f = **d** + **n =** g**c** + **v** + **n.**

Finally, VIF is computed as the ratio of the mutual information between **c** and **f** and the mutual information between **c** and **e** for all sub-bands except the approximation sub-band. The mutual information is conditioned on knowledge of estimated scalar multiplier z:

$$\mathrm{I}(\mathbf{c};\mathbf{e} \mid z) = \frac{1}{2}\log\frac{\left|z^2\mathbf{C_u} + \sigma_n^2\mathbf{I}\right|}{\left|\sigma_n^2\mathbf{I}\right|} = \frac{1}{2}\sum_{j=1}^{M}\log\left(1 + \frac{z^2\lambda_j}{\sigma_n^2}\right);$$

$$\mathrm{I}(\mathbf{c};\mathbf{f} \mid z) = \frac{1}{2}\log\frac{\left|g^2z^2\mathbf{C_u} + (\sigma_v^2 + \sigma_n^2)\mathbf{I}\right|}{\left|(\sigma_v^2 + \sigma_n^2)\mathbf{I}\right|} = \frac{1}{2}\sum_{j=1}^{M}\log\left(1 + \frac{g^2z^2\lambda_j}{\sigma_v^2 + \sigma_n^2}\right).$$

The covariance matrix is factorized as $\mathbf{C}_u = \mathbf{Q}\Lambda\mathbf{Q}^T$, where Λ is a diagonal matrix whose diagonal entries are the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_M$. The overall VIF Index is then

$$\mathrm{VIF} = \frac{\mathrm{I}(\mathbf{C};\mathbf{F} \mid \mathbf{z})}{\mathrm{I}(\mathbf{C};\mathbf{E} \mid \mathbf{z})} = \frac{\sum_{i=1}^{N}\mathrm{I}(\mathbf{c}_i;\mathbf{f}_i \mid z_i)}{\sum_{i=1}^{N}\mathrm{I}(\mathbf{c}_i;\mathbf{e}_i \mid z_i)},$$

where i is the index of local coefficient patches, with all N sub-bands included.

To illustrate the performance of the VIF Index, refer again to Fig. 1. The scores are again very much in line with subjectivity. Again, what matters is MOS. SSIM, VIF and most other well-known IQA algorithms were evaluated in a massive study (consisting of > 750 images and more than 25,000 subjective judgments) as reported in "An Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Trans. on Image Processing,* vol. 15, no. 11, pp. 3440-3451, Nov. 2006. This conclusive study showed SSIM and VIF to perform comparably, but far better than all prior algorithms.

3. Video Quality Assessment

Most existing VQA algorithms are derived from IQA algorithms. Some apply still IQA algorithms frame by frame on a video sequence. The MSE/PSNR is such a metric that remains heavily used as a video quality metric, due to its simplicity and ease of use. Simple versions of SSIM have been used for VQA, and shown to be quite competitive with other VQA algorithms. Some VQA algorithms are derived from IQA algorithms by incorporating a temporal filtering block. Temporal filtering has also been used to extend the SSIM and VIF metric to the video domain. Yet, correlation of the best VQA algorithms with subjectivity is *much* lower than for IQA algorithms.

No existing VQA algorithms attempt to use motion information directly. Yet motion plays a crucial role in perception of videos and needs to be accounted for. The HVS is quite sensitive to motion and can accurately judge the velocity and direction of moving objects. Moving objects attract our attention and play an important role in visual salience. Indeed considerable resources are devoted to motion perception in the HVS. Hence, simple extensions of IQA techniques to video are unlikely to correlate well with perceptual quality.

We are trying to motivate a new paradigm for VQA that incorporates motion modeling. Some spatial artifacts occur within video frames that do not arise from temporal processes, such as blocking from DCT coefficient quantization in MPEG and H.264; ringing from quantization in block-free codecs such as Motion JPEG-2000; mosaic patterns; and false contouring. Spatio-temporal artifacts arise from spatio-temporal processes including ghosting behind fast-moving objects; block artifacts from faulty motion compensation; mosquito effect near moving edges; jerkiness from temporal aliasing or transmission delays; and smearing from slow acquisition. Although existing algorithms are successful in detecting spatial distortions, they usually fail to adequately capture the temporal distortions.

We are attempting to ameliorate this by incorporating motion information in video sequences. We outline the new Video Structural SIMilarity (V-SSIM) Index (see "A Structural Similarity Metric for Video Based on Motion Models," *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, April 2007). In V-SSIM, short video segments are modeled as translating image patches. Under this assumption, the Fourier transform of the video patch lies in a plane in the frequency domain. The orientation of this plane is determined by the speed and direction of translation, while the frequencies in the plane are identical to the spatial frequencies contained in the image patch undergoing translation. Thus, image motion takes an elegant and accessible form. Decomposition using a sub-band family captures this form and facilitates VQA. Similar decompositions are believed to occur in the HVS. For

these reasons, the V-SSIM index is computed using the outputs of a sub-band filter family operating on reference and test videos. Decomposition using a sub-band family allows motion estimation and quality computation as described below.

V-SSIM can be described as shown in Fig. 2. Reference and test signals are decomposed by a family of Gabor filters forming band-pass spatio-temporal frequency channels. The center frequencies of all Gabor filters are at a fixed radius from the origin, lying on a sphere. Iso-surface contours of each filter are shown in Fig. 2. The sub-band outputs on the reference are used to compute motion estimates. At each pixel, a 2-D vector that specifies the speed of movement of the pixel is obtained. Motion computation is performed using the Fleet and Jepson algorithm. The motion vector at each pixel is used to select a subset of the filter family lying in close proximity to the plane containing the local reference spectrum. We require the local spectral plane to lie within one standard deviation of the Gabor center frequency (Fig. 2). The V-SSIM Index at each pixel is then computed as the SSIM Index between the selected subset of sub-band coefficients of the reference and test video at that pixel. This filter selection rule is a form of motion compensated filtering that allows V-SSIM to capture spatial, temporal, and spatio-temporal distortions by computing quality along the motion trajectories of the reference video.

The V-SSIM Indices are displayed as a quality map in Fig. 2 that pinpoints those video regions that suffer from poor quality. The V-SSIM Index of the entire video is obtained as the mean of this quality map. The V-SSIM algorithm was tested on a database created by the Video Quality Experts Group (VQEG) containing reference and distorted videos as well as quality scores assigned by human observers. The results are quite promising since the V-SSIM outperforms prior methods including frame-by-frame SSIM. These results validate our claims regarding the importance of modeling motion and temporal artifacts in VQA.

4. Summary

We discussed recent objective algorithms that compute the perceptual quality of digital images or video sequences. We focused our discussion on two recent algorithms for IQA that represent the state-of-the-art, the SSIM and VIF quality indices. We also discussed the recent V-SSIM quality index and showed that incorporating motion models can result in significant gains in the performance of VQA algorithms.

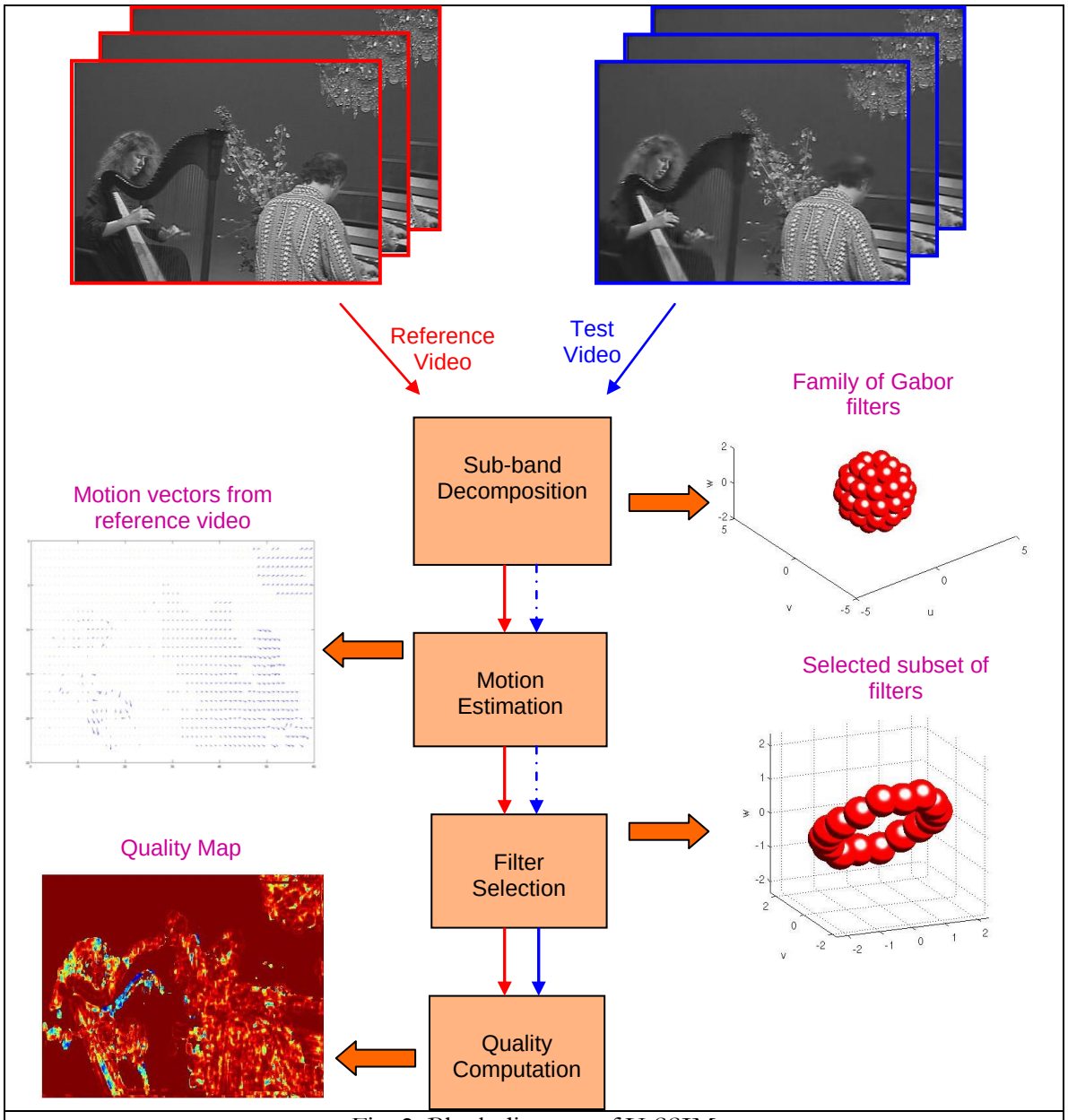


Fig. 2: Block diagram of V-SSIM