A STRUCTURAL SIMILARITY METRIC FOR VIDEO BASED ON MOTION MODELS

Kalpana Seshadrinathan and Alan C. Bovik

Department of Electrical and Computer Engineering,
1, University Station C0803, The University of Texas at Austin, Austin, TX - 78712

ABSTRACT

Quality assessment plays a very important role in almost all aspects of multimedia signal processing such as acquisition, coding, display, processing etc. Several objective quality metrics have been proposed for images, but video quality assessment has received relatively little attention and most video quality metrics have been simple extension of metrics for images. In this paper, we propose a novel quality metric for video sequences that utilizes motion information in video sequences, which is the main difference in moving from images to video. This metric is capable of capturing temporal artifacts in video sequences in addition to spatial distortions. Results are presented that demonstrate the efficacy of our quality metric by comparing model performance against subjective scores on the database developed by the Video Quality Experts Group.

Index Terms— Quality Assessment, Video Signal Processing, motion compensation, Video Quality Experts Group (VQEG)

1. INTRODUCTION

With the rapid increase in popularity of multimedia applications such as Video On Demand, wireless video, digital cinema etc., it is natural that the question of video quality control become a central concern. Unfortunately, advances in video processing and communication have not been matched by similar progress in methods for video performance and quality analysis. Unlike many signal processing applications, the intended receiver of the video signal is nearly always the human eye-brain, which remains only weakly modeled. Video Quality Assessment (VQA) algorithms attempt to assess *perceptual degradations* introduced by video acquisition, processing and communication devices. Although progress in the development of accurate and reliable VQA algorithms has been slow, great strides have recently been made in assessing the quality of still images [1, 2]. In this paper, we develop a full reference quality metric for video signals by making natural extensions of the powerful Structural SIMilarity (SSIM) metric for still images to the spatio-temporal (video) domain. Full reference quality metrics assume the availability of a "perfect" reference video and attempt to assess the fidelity of the test video with respect to this pristine original.

Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) remain heavily used as video quality metrics, despite their poor correlation with visual quality, due to their simplicity and the lack of a reliable alternative [2]. Most of the research on VQA over the past twenty years has focused on methods that attempt to model the Human Visual System (HVS). The premise behind such HVS-based metrics is to process the visual data by simulating the visual pathway of the eye-brain system. Examples of video quality metrics based on the HVS-based philosophy include the Digital Video Quality (DVQ) metric [3], the Sarnoff JND model [4]and the Perceptual Distortion Model (PDM) [5]. However, studies conducted by the Video Quality Experts Group indicate that the performance of HVS-based VQA algorithms leaves considerable room for improvement [6]. HVS-based VQA metrics suffer from inaccurate modeling of the HVS and in particular, temporal mechanisms in the HVS is a likely source of performance loss as well. For example, 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 the visual cortex only and these models are too simple to describe motion processing in the HVS. In particular, activity of neurons in area MT of the extra-striate cortex, which play a very important role in motion perception, is not accounted for in any of these models.

Very simple and preliminary extensions of the SSIM index have been proposed for VQA [7] using a simple frame-by-frame implementation of the SSIM image quality metric. However, this metric does not utilize *motion information* or model temporal artifacts in video that can affect the quality of the video sequence. The human eye is quite sensitive to motion and can accurately judge the velocity and direction of moving objects - unsurprising given the relevance of these skills to survival. Considerable resources in the HVS are devoted to motion perception and it is hence essential for video quality metrics to incorporate some form of motion modeling. Further, video sequences suffer from *spatio-temporal* artifacts and frame-by-frame quality metrics cannot account for temporal distortions in videos. Example of such temporal artifacts include ghosting, jitter, motion compensation mismatch, smearing, mosquito noise etc.

We believe that the performance of video quality assessment 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 which greatly motivates our work. In this paper, we present a Video Structural SIMilarity index, known as V-SSIM, that incorporates motion modeling using optical flow. This results in a *motion compensated* implementation of the structural similarity metric. We then demonstrate the efficacy of our metric on the VQEG database that contains distorted sequences as well as subjective scores assigned by human observers to these sequences [6].

2. V-SSIM INDEX FOR VIDEO SEQUENCES

2.1. Motion in the frequency domain

In this paper, we consider the apparent motion of image intensities, namely the *optical flow*. The term velocity denotes the optical flow vector and not true three dimensional velocity of motion. Let $i(x, y)$ denote an image and let $\tilde{I}(w_x, w_y)$ denote its Fourier transform. Assuming that this image undergoes translation with a velocity $\vec{v} = (v_x, v_y)$, the resulting video sequence is given by $f(x, y, t) = i(x - v_x t, y - v_y t)$. Then, $\tilde{F}(w_x, w_y, w_t)$, the Fourier transform of $f(x, y, t)$, lies entirely along a plane in the frequency

domain [8]. This plane is defined by:

$$v_x w_x + v_y w_y + w_t = 0$$

Additionally, the magnitudes of the spatial frequencies do not change but are simply sheared in the frequency domain. It can be shown that $\tilde{F}(w_x, w_y, w_t)$ is given by

$$\tilde{F}(w_x, w_y, w_t) = \begin{cases} \tilde{I}(w_x, w_y) & \text{if } v_x w_x + v_y w_y + w_t = 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

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 can be used *locally* to describe video sequences, since translation is a linear approximation to more complex types of motion. Eq. (1) provides us with an explicit characterization of the motion of a video sequence in the frequency domain. Frequency domain approaches are also well suited to our study of human perception of video signals due to the presence of bandpass visual channels in the HVS [5]. Hence, in the proposed V-SSIM video quality assessment system, the video sequence is filtered spatio-temporally using a family of band-pass filters and quality assessment is performed on the resulting bandpass channels in the spatio-temporal frequency domain.

2.2. SSIM index for images

Although the SSIM index was initially proposed in the pixel domain, a complex wavelet version was proposed in [9] and the design of our metric closely follows the development of the wavelet domain version. Hence,we briefly overview the Complex Wavelet Structural SIMilarity (CW-SSIM) index for images. The reference and test images are filtered using a family of complex wavelets consisting of N filters. Let $\vec{R} = \{r_k, k = 1, 2, \ldots, N\}$ and $\vec{S} = \{s_k, k = 1, 2, \ldots, N\}$ denote a set of coefficients from the reference and distorted images at corresponding spatial locations. Then, the CW-SSIM index between these coefficients is given by

$$\text{CW-SSIM}(\vec{S}, \vec{R}) = \frac{2\sum_{k=1}^{N} r_k s_k^* + K}{\sum_{k=1}^{N} |r_k|^2 + \sum_{k=1}^{N} |s_k|^2 + K} \quad (2)$$

where c^* denotes the complex conjugate of c, $|c|$ denotes the magnitude of c and K is a small positive constant added to prevent prevent numerical instability when the value of the denominator is very low. The overall quality index of the entire image is then calculated as the mean of the CW-SSIM indices over all the pixels of the image. This quality measure was shown to perform very well in predicting the quality of still images [9, 10].

2.3. Selection of sub-band filter family

In Section 2.1, we discussed the simple form that motion in video sequences takes in the frequency domain. This motivates us to perform our analysis in the frequency domain. Therefore, we will perform a decomposition of the video sequence into bandpass channels in the frequency domain and this decomposition helps us achieve two goals. Firstly, optical flow estimation can be performed using the outputs of these bandpass channels. Secondly, similar to the CW-SSIM index, our proposed video quality metric will compute similarity indices between these bandpass filtered outputs in the frequency domain, as opposed to the pixel domain.

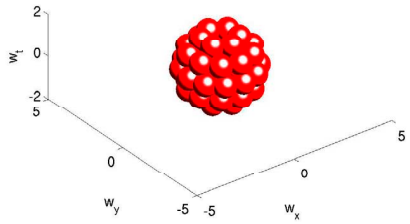


Fig. 1. Geometry of the Gabor filterbank in the frequency domain.

Although any filter family can be used to decompose the video sequence into bandpass channels, we opt to use Gabor filters in our implementation. Evidence indicates that the receptive field profiles of simple cells in the mammalian visual cortex can be described well by a set of Gabor filters [11]. Also, Gabor filters attain the theoretical lower bound on the uncertainty in the frequency and spatial variables and thus, visual neurons can be said to optimize the uncertainty in information resolution [11]. Additionally, development of the video quality metric in Section 2.5 requires estimation of the optical flow vectors and Gabor filters have been successfully used for this purpose in the literature [12].

A Gabor filter $g(x, y, t)$ is simply the product of a Gaussian window and a complex exponential and is given by:

$$g(x, y, t) = \frac{1}{(\sqrt{2\pi})^3 \sigma_x \sigma_y \sigma_t} e^{-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2}\right)} e^{i(Ux+Vy+Wt)} \quad (3)$$

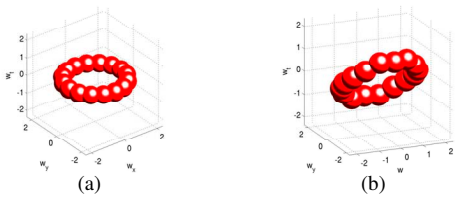
where (U, V, W) is the center frequency of the Gabor filter and $(\sigma_x, \sigma_y, \sigma_t)$ is the spread of the Gaussian window in space-time. Then, the Fourier transform of the Gabor filter is a Gaussian whose standard deviation in the frequency domain is $(1/\sigma_x, 1/\sigma_y, 1/\sigma_t)$ and is given by:

$$\tilde{G}(w_x, w_y, w_t) = e^{-\frac{1}{2}\left[\sigma_x^2 (w_x - U)^2 + \sigma_y^2 (w_y - V)^2 + \sigma_t^2 (w_t - W)^2\right]} \quad (4)$$

The filters that we used in our implementation have the same geometry as the Gabor filters described in [12] and are illustrated in Figure 1. We used a family of filters consisting of $N = 22$ filters all at the same scale, i.e., all filters are tuned to the same spatio-temporal frequency band. Figure 1 shows isosurface contours of the resulting filter bank in the frequency domain. We used filters with rotational symmetry and the spatial spread of the Gaussian filters is the same along all axes.

2.4. Optical flow estimation

The proposed V-SSIM algorithm uses motion information from the reference video sequence in the form of the optical flow vector and we briefly describe the optical flow estimation algorithm. We used the Fleet and Jepson phase based algorithm for optical flow estimation with slight modifications [12]. This algorithm attempts to find constant phase contours of the outputs of a Gabor filterbank to estimate the optical flow vectors. Constant phase contours are computed by estimating the derivative of the phase of the Gabor filter outputs, which in turn can be expressed as a function of the derivative of the Gabor filter outputs. The algorithm in [12] uses a 5-point central



(a) (b)

Fig. 2. Illustration of a set of motion compensated filters: (a) A static sequence (b) Sequence with motion.

difference kernel to perform the derivative computation. However, we chose to perform the derivative computation by convolving the video sequence with filters that are derivatives of the Gabor kernels denoted by $g'_x(x,y,t), g'_y(x,y,t), g'_y(x,y,t)$.

$$g'_x(x,y,t) = g(x,y,t)\left(\frac{-x}{\sigma_x^2} + iU\right) \tag{5}$$

Similar definitions apply for the derivatives along y and t directions. This filter is more accurate in computing the derivative of the Gabor outputs and produced better optical flow estimates in our experiments. We wish to point out that the Fleet and Jepson algorithm does not produce flow estimates with 100% density, i.e. flow estimates are not computed at each and every pixel of the video sequence. Instead, optical flow is only computed at pixels where there is sufficient information to do so. Thus, flow is not computed in regions of the video sequence that suffer from the aperture problem, or areas whose frequency components lie significantly outside the bandpass region of the Gabor filters. However, this is not a serious issue since we compute spatial quality indices at these locations as described in Section 2.5. Finally, note that our current implementation uses only one scale of filters and cannot compute optical flow in fast moving regions of the video sequence due to temporal aliasing [12]. We are working on a multi-scale framework for flow computation.

2.5. Proposed quality index for video sequences

Motion plays a key role in the perception of video sequences and distorted videos suffer from artifacts that are *spatio-temporal* as described in Section 1. We hence made a case for the importance of modeling of motion as well as temporal distortions in a video quality assessment system. In Section 2.4, we described a method to estimate the motion in a video sequence that has been proposed in the literature. We are now in a position to use these optical flow estimates to derive a structural similarity index for video sequences.

Let $\{\vec{R} = r_k, k = 1, 2, \ldots N\}$ denote coefficients at a pixel obtained by filtering the reference video sequence with the Gabor filter family $\{g_k(x,y,t), k = 1, 2, \ldots N\}$. Similarly, let $\{\vec{S} = s_k, k = 1, 2, \ldots N\}$ denote coefficients at the corresponding spatio-temporal location obtained by filtering the distorted sequence, whose quality we wish to estimate, with the Gabor filterbank.

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 then identify the Gabor filters that overlap significantly with this plane. In our implementation, we required that the plane lie within one standard deviation of the Gabor filter in the frequency domain. Thus, if the optical flow vector at a pixel is (v_x, v_y) and the center frequency of the Gabor filter is (U_k, V_k, W_k), then the plane that contains the spectrum of the video sequence is described by $v_x w_x + v_y w_t + w_t = 0$. Thus, our rule for selection of the filter would be defined by:

$$C = \left\{ k : \left| \frac{v_x U_k + v_y V_k + W_k}{\sqrt{v_x^2 + v_y^2 + 1}} \right| \leq \frac{1}{\sigma} \right\} \tag{6}$$

where C denotes a set that contains the selected filter indices and σ denotes the standard deviation of the Gabor filter along any axis in the space domain. Selection of such a subset of filters results in *motion compensated* filtering of the video sequence. However, our flow estimation algorithm does not produce flow estimates at each pixel of the video sequence. At pixels without motion information, we simply set $v_x = v_y = 0$. This results in the computation of V-SSIM indices that capture spatial distortions alone at these pixels.

The V-SSIM index that we now propose closely resembles the CW-SSIM index between the outputs of the Gabor filters $\vec{R}$ and $\vec{S}$ for images. However, we compute the V-SSIM index *only* between the outputs of those filters that satisfy Eq. (6) using:

$$\text{V-SSIM}(\vec{R}, \vec{S}) = \frac{2\sum_{k\in C} |r_k||s_k| + K}{\sum_{k\in C} |r_k|^2 + \sum_{k\in C} |s_k|^2 + K} \tag{7}$$

Note that we only use the magnitudes of the Gabor filter outputs to compute the V-SSIM index and contrast this with the definition of the CW-SSIM index in Eq. (2). The reason for computing CW-SSIM using the complex wavelet response was to design a translation insensitive measure, and the phase of the complex wavelet response corresponds to small translations in the image. However, in the video scenario, the phase of the Gabor outputs represent *motion information* and the Fleet and Jepson optical flow estimation algorithm computes flow using this phase information. Thus, once motion compensation has been accomplished, we compute the V-SSIM index only between the magnitudes of the filter outputs.

We hypothesize that our proposed metric is capable of handling a wide variety of both spatial as well as temporal artifacts and will now attempt to provide some insight on this. Consider the case of distortions that are entirely spatial, where the local orientation of the plane is identical in both the reference as well as the distorted sequence. In this situation, once the Gabor filters that intersect the plane have been identified, the proposed V-SSIM index is simply a motion compensated implementation of the CW-SSIM index. As an example, consider a video sequence that consists of the same image repeated over frames. In this situation, the entire frequency spectrum of the plane lies along $w_t = 0$ and all spatial filters with small temporal frequency component will intersect this plane. Thus, the proposed method is equivalent to computing the CW-SSIM index of the image with a set of Gabor filters. This is illustrated in Fig. 2(a). When the image sequence undergoes translational motion, the filters closest to the plane containing the frequency spectrum of the image are identified and this corresponds to motion compensated filtering of the video sequence. This is illustrated in Fig. 2(b). If the motion in the reference sequence is accurately represented in the distorted sequence as well, the V-SSIM index computed using the motion compensated filters will correspond to the spatial similarity index between the sequences. In this situation, the video only suffers from spatial artifacts.

Now, consider the situation where there are certain temporal distortions in the video. Quantization of the motion vectors, ghosting or motion compensation mismatch would correspond to a misalignment of the planes containing the frequency spectrum of the reference and distorted sequences. This is illustrated in Fig. 3. Thus,

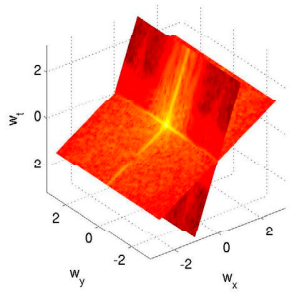


Fig. 3. Illustration of temporal artifacts in the frequency domain.

computing the V-SSIM index between these sequences would lead to smaller similarity indices as compared to the case where there are no motion artifacts. Also, small misalignments are not likely to produce large differences in the filter outputs, while large changes which may occur due to, for example, temporal aliasing in the video will lead to a very poor similarity index between the videos.

3. RESULTS

We tested our proposed V-SSIM index on the VQEG database [6]. This 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. As mentioned in Section 2.4, 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. These excluded sequences were sequence 6 (Formula 1 racing car), sequence 8 (scrolling text), sequence 9 (rugby game) and sequence 19 (football game). All the VQEG test sequences are interlaced and our algorithm operates only on the odd fields of the interlaced sequences. To reduce computational burden, flow and V-SSIM indices were not computed for each frame, but only for one in 16 frames. Also, Eq. (7) computes the VSSIM index at each pixel only using a subset of the filtered outputs at that pixel. In our implementation, we computed the VSSIM index at a pixel using the selected filter outputs from a 5×5 spatial window centered around the pixel.

The results of our simulations on the remaining 16 reference sequences with 288 data points is summarized in Table 1, which shows the Spearman Rank Order Correlation Coefficient (SROCC) between subjective and objective scores for different video quality metrics. SROCC is one of the metrics specified by the VQEG that tests the prediction monotonicity of a video quality assessment system. PSNR does not correlate well with subjective scores as seen in Table 1. Proponent P8 is the best performing metric amongst the 10 different proponent models tested by the VQEG in terms of the SROCC metric [6]. We also compare our results against the better performing version of the two metrics proposed in [7]. The results clearly indicate that our V-SSIM index performs very well and is competitive with other video quality assessment systems. In fact, the proposed metric out-performs all the metrics that we compared against in prediction monotonicity.

Prediction Model	SROCC
Peak Signal to Noise Ratio	0.786
Proponent P8 (Swisscom)	0.803
Metric in [7]	0.812
Proposed V-SSIM	0.835

Table 1. Comparison of SROCC values for different video quality assessment algorithms.

4. CONCLUSIONS AND FUTURE WORK

In conclusion, we proposed a novel framework for the quality assessment of video sequences, that incorporates explicit modeling of motion and captures spatial as well as temporal artifacts in video sequences. In the future, we would like to develop a multi-scale framework for optical flow estimation which would enable us to test our algorithm on fast moving video sequences as well.

5. REFERENCES

[1] Z. Wang and A. C. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE*, vol. 9, no. 3, pp. 81–84, 2002.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[3] 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.

[4] (2003) Sarnoff corporation, JNDMetrix Technology. [Online]. Available: http://www.sarnoff.com/products services/video vision/ jndmetrix/downloads.asp

[5] 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, May 1999, pp. 175–184.

[6] (2000) Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[7] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[8] A. B. Watson and J. Ahumada, A. J., "Model of human visual-motion sensing," *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 2, no. 2, pp. 322–342, 1985.

[9] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 2, 2005, pp. 573–576.

[10] Z. Wang and A. C. Bovik, *Image Quality Assessment*. New York: Morgan and Claypool Publishing Co., 2006.

[11] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 2, no. 7, pp. 1160–1169, 1985.

[12] D. Fleet and A. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, 1990.