STATISTICAL VIDEO MODELS AND THEIR APPLICATION TO QUALITY ASSESSMENT

Kalpana Seshadrinathan and Alan C. Bovik

Dept. of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX - USA.

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. Most of the video quality metrics in the literature are simple extensions of metrics for images.

In this paper, we integrate natural image statistics and the theory of optical flow to propose a new model for the statistics of video signals in the wavelet domain. This model utilizes motion information in video sequences, which is the main difference in moving from images to video. Results are presented to demonstrate the effectiveness of this model to describe the statistics of wavelet coefficients. We then briefly describe how this model can be used in an information theoretic framework to develop quality metrics for video sequences.

1. INTRODUCTION

Accurate objective quality metrics are of great potential benefit to the video industry, as they promise the means to evaluate the performance of acquisition, display, coding and communication systems. Although a lot of work has been done on still image quality assessment, surprisingly little work has been done on quality assessment of video signals. Even today, mathematical measures such as the Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) are widely used in tasks such as the design of image communication systems, although it is well known that these metrics don't correlate well with *visual quality*. The popularity of PSNR is partly due to its mathematical convenience and simplicity and partly due to the lack of a competing metric that has been shown to consistently perform better in predicting visual quality, across images.

Many of the proposed quality metrics in the literature use models that describe the frequency response, luminance and contrast sensitivities, contrast masking and other features of the Human Visual System (HVS) [1]. Visual quality is then computed as the distance between the distorted and reference images, after normalizing these signals to account for the sensitivities of the HVS. However, the performance of these metrics is limited by the accuracy and complexity of the underlying HVS models. More recently, the SSIM or Wang-Bovik index and VIF or Sheikh-Bovik index have been shown to be highly effective in predicting image quality and extensions to evaluating video quality have been proposed [2, 3]. However, neither of these metrics attempt to model motion in video sequences.

Biological vision systems devote considerable resources to motion processing, since estimation of the speed and direction of motion of objects in the environment are crucial to the survival of the organism. Presentation of video sequences to human subjects induces visual experience of motion and perceived distortion in video sequences is a combination of both spatial and motion artifacts. For example, motion artifacts such as ghosting and blocking are clearly visible in video signals distorted by compression, blurring etc. Thus, video quality assessment is not a straight forward extension of image quality assessment. Modeling motion and distortions in motion is essential in the development of a video quality metric, and optical flow is a valuable tool that describes the apparent motion of image intensities. In this paper, we propose a new model to describe the statistics of three-dimensional wavelet coefficients of video sequences as a function of optical flow. We then briefly describe how these statistical models can be used to develop a video quality metric in an information theoretic framework.

2. STATISTICAL MODEL FOR VIDEO

In this section, we present a novel model that describes the statistics of the three-dimensional wavelet coefficients of video signals. In Section 2.1, we review how translational motion in video manifests itself in the frequency domain. In Section 2.2, we propose a model to describe the statistics of two-dimensional scenes in video sequences in the frequency domain. In Section 2.3, we consider motion of these scenes and thereby derive the statistics of the wavelet coefficients of video signals. Finally, Section 2.4 describes how the parameters in the model can be estimated.

Supported by the National Science Foundation

2.1. Motion in the Frequency Domain

In this paper, we only consider apparent motion of image intensities, namely the *optical flow*, and 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 $f(x, y, t) = i(x - v_x t, y - v_y t)$. If $\tilde{F}(w_x, w_y, w_t)$ denotes the Fourier transform of $f(x, y, t)$, then $\tilde{F}(w_x, w_y, w_t)$ lies entirely along a plane in the frequency domain: $v_x w_x + v_y w_y + w_t = 0$ [4]. 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

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

We assume that short segments of video, without any scene changes, consist of image patches undergoing translation. This model can be used to *locally* describe video sequences, since translation is a linear approximation to more complex types of motion.

2.2. Modeling the Statistics of Images

The Gaussian Scale Mixture (GSM) model is a popular approach for describing the statistics of wavelet coefficients of natural images [5]. Wavelet coefficients of natural images are not modeled well by independent and identically distributed (i.i.d.) Gaussian random variables, a model that is often used due to its mathematical tractability. Wavelet coefficients at adjacent positions, scales and orientations tend to have similar magnitudes, due to the presence of oriented structures in images such as edges. GSM random variables have been used to model the statistics of wavelet coefficients successfully. A random vector is said to be a GSM if it is a product of a scalar random variable, known as the mixing density, and a Gaussian random vector. Here, the mixing density models the dependencies between neighboring wavelet coefficients.

Edges, occlusion boundaries and other oriented structures in natural images manifest as oriented components with large magnitudes across scales in the Fourier transform of the image. Image texture is characterized by a high concentration of localized spatial frequencies in the frequency domain. Due to these reasons, localized regions in the frequency domain representation of natural images tend to have similar magnitude and can hence be modeled well using a GSM in *continuous* frequency space.

We propose a continuous GSM random field model for the Fourier transform of a natural image that is restricted to a small region of the frequency domain corresponding to a sub-band of a scale-space decomposition:

$$\tilde{I}(w_x, w_y) \sim zU(w_x, w_y), \quad (w_x, w_y) \in \mathrm{S}$$

where S denotes a specific sub-band, z is the multiplier or mixing density and U is a complex, zero-mean, white Gaussian random field. The z field has been modeled using a gamma density in the literature [5], but in this paper, we assume that it is a constant parameter and do not assume any prior knowledge of the distribution of z. We denote this estimated value of z by $\hat{z}$ and the estimation details are presented in Section 2.4.

2.3. Incorporating Motion Models

We noted in Section 2.1 that when an image moves at a velocity $\vec{v}$, the frequency spectrum of this image sequence is simply the Fourier transform of the image, but sheared at an orientation defined by the velocity vector. Using the model proposed in Section 2.2 for the image, we have the following distribution for each subband in the frequency spectrum of the video sequence:

$$\tilde{F}(w_x, w_y, w_t) \sim zU(w_x, w_y)\delta(v_x w_x + v_y w_y + w_t)$$

where $\delta(t)$ is the Dirac delta function.

Consider filtering this video signal with a family of three-dimensional sub-band filters. Although any filter family can be used, we opt to use Gabor filters in our analysis. 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 [6]. 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 [6]. Gabor filters are hence highly suitable for use in video quality assessment which deals with human perception of video sequences. Additionally, development of the video quality metric in Section 4 requires estimation of the optical flow vectors and Gabor filters have been successfully used for this purpose in the literature [7].

Let $g(x, y, t)$ denote a Gabor filter and let $\tilde{G}(w_x, w_y, w_t)$ denote its Fourier transform. Consider the wavelet coefficients in a particular sub-band, denoted by $w(x, y, t)$, obtained by filtering the video signal $f(x, y, t)$ with the Gabor filter $g(x, y, t)$. We then have

$$w(x, y, t) = \int g(x', y', t') f(x - x', y - y', t - t') \mathrm{d}x' \mathrm{d}y' \mathrm{d}t' \tag{1}$$

$$= \frac{1}{8\pi^3} \int \tilde{G}(w_x, w_y, w_t) \tilde{F}(w_x, w_y, w_t) \mathrm{d}w_x \mathrm{d}w_y \mathrm{d}w_t \tag{2}$$

Eq. (2) is a consequence of the properties of the Fourier transform and Parseval's theorem. We assume $f(x', y', t')$ denotes the video signal centred at (x, y, t) and hence ignore the phase shift term that would have appeared in Eq. (2).

Given that the mixing density is known, $\tilde{F}(w_x, w_y, w_t)$ has been modeled as a Gaussian random field. We use the estimated value of the mixing field, namely $\hat{z}$, in our analysis here. Since $w(x, y, t)$ as defined in Eq. (2) is the integral of a linear function of a Gaussian random process, it is a Gaussian random variable. It can then be shown that $w(x, y, t)$ has zero-mean and variance σ_w^2 given by:

$$\sigma_w^2 = \left(\frac{1}{8\pi^3}\right)^2 \int \hat{z}^2 |\tilde{G}(w_x, w_y, -v_x w_x - v_y w_y)|^2 \mathrm{d}w_x \mathrm{d}w_y \tag{3}$$

Here, $\tilde{G}(w_x, w_y, -v_x w_x - v_y w_y)$ is a two-dimensional slice of the Gabor filter along the plane containing the frequency spectrum of the translating video signal. The variance of the wavelet coefficients is hence a function of the energy of the Gabor filter along this plane. This makes intuitive sense, since only those filters that intersect the oriented plane will produce large magnitude coefficients. σ_w^2 is also a function of $\hat{z}^2$, which models the average energy of the image $i(x, y)$ in the spatial frequency band spanned by the filter. This also agrees with our intuition, since the magnitude of the wavelet coefficients will also depend on the magnitude of $\tilde{F}(w_x, w_y, w_t)$ along the oriented plane. Eq. (3) explicitly characterizes the distribution of the wavelet coefficients as a function of the optical flow vector and can be evaluated in closed form.

2.4. Parameter Estimation

The value of the mixing density, namely $\hat{z}$, needs to be estimated from the given video sequence. Denoting the first frame of the video sequence by $i(x, y)$, we need to estimate the energy of this image in the sub-bands spanned by the Gabor filters. Note that $\hat{z}$ is the energy of the image in the two-dimensional sub-band obtained by the projection of the Gabor filter onto the plane $w_t = 0$. This can be estimated by filtering $i(x, y)$ using a family of two-dimensional Gabor filters corresponding to these projections. $\hat{z}$ can them be computed *locally* as the energy in these filtered signals.

3. RESULTS

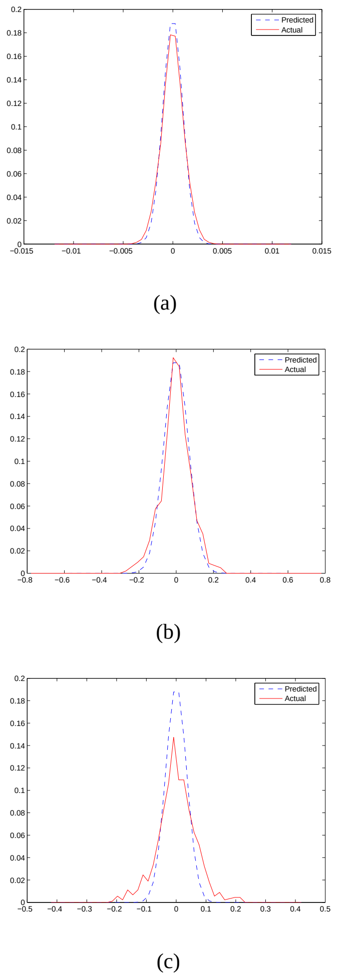
To test our model, we implemented a family of sine phase Gabor filters and used these to filter sequences with known optical flow vectors. Letting w denote the wavelet coefficient at a specific spatiotemporal location, our model states that $w/\hat{z}$ is normally distributed with zero mean and variance $\sigma_w^2/\hat{z}^2$. We plotted the distribution of $w/\hat{z}$ against this predicted distribution. Figure 1(a) shows the result on a video sequence consisting of a repeated image, *i.e.*, there is no motion in the entire sequence. The fit is quite good and similar plots were obtained for all images in our database. We also tested our model on the Yosemite fly through sequence [8]. Since the predicted variance σ_w^2 is a function of v_x and v_y, the velocity vectors were quantized and distributions were plotted for pixels that have the same velocity. Figures 1(b) and 1(c) show the results obtained for different values of the optical flow vector [8]. These results show that our model performs quite well in predicting the statistics of three-dimensional video wavelet coefficients.

4. VIDEO QUALITY METRIC

Recently, researchers have proposed information theoretic approaches to the image quality assessment problem [9]. In this framework, a natural image source is assumed to transmit reference images over a communication channel. The communication channel is used to model the distortions that the reference image undergoes and the test image is assumed to be the output of this channel. The mutual information between the image source and the output of this channel is then used to quantify the quality of the distorted image. The success of this approach in image quality prediction can be attributed to the accurate models used to describe the statistics of the natural image source [10]. It has been argued that natural scene and HVS modeling are dual problems, as the HVS has evolved in response to viewing natural scenes [11]. Thus, source models provide a new perspective on the quality assessment problem, which has traditionally been attacked using HVS models. We use the statistical model for wavelet coefficients proposed in Section 2 to derive a quality metric for video signals, closely following the development in [9] for images.

In [9], a simple attenuation and additive noise distortion model in the wavelet domain is used for images. The blur component models any loss of signal energy in wavelet sub-bands, due to operations such as compression, blurring etc. This model, when used *locally*, has been shown to be adequate in modeling several distortions like compression, additive noise, blurring and contrast stretching [9].

We suggest that a similar model can be used to describe most motion artifacts such as ghosting and blocking. In this case, the distortion in the optical flow vector of the test video is modeled using a blur and additive noise model. This is illustrated in Fig. 2. Let X_d denote the wavelet coefficients of the distorted video. The quality of the distorted signal can then be quantified by the mutual information between the input and output of this channel, namely $I(X; X_d)$. The proposed statistical model can then be used to compute this quantity in closed form.



(a)

(b)

(c)

Fig. 1. Distribution of the normalized wavelet coefficients $w/\hat{z}$. Dashed line shows the predicted distribution.

5. CONCLUSIONS AND FUTURE WORK

In conclusion, we propose a novel statistical model to describe the wavelet coefficients of video signals. We also propose a framework to predict the visual quality of video signals using this model. Experimental studies indicate that our model is successful in predicting the distribution of wavelet coefficients of video sequences. These results are promising and we envision that the proposed metric that we are developing will also be successful in predicting visual quality.

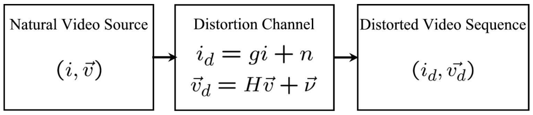


Fig. 2. Block diagram of proposed video quality metric

6. REFERENCES

[1] A. B. Watson, J. Hu, and I. J.F. Mc Gowan., “Dvq: A digital video quality metric based on human vision,” *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20–29, 2001.

[2] Z. Wang, L. Lu, and A. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Communication, special issue on objective video quality metrics*, vol. 19, Jan.

[3] H. R. Sheikh and A. C. Bovik, “A visual information fidelity approach to video quality assessment,” in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005.

[4] A. B. Watson and A. J. Ahumada, “Model of human visual motion sensing,” *Journal of the Optical Society of America*, vol. 2, pp. 322–342, 1985.

[5] M. J. Wainwright and E. P. Simoncelli, “Scale mixtures of Gaussians and the statistics of natural images,” in *Adv. Neural Information Processing Systems (NIPS*99)*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds., vol. 12. Cambridge, MA: MIT Press, 2000, pp. 855–861.

[6] J. 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 2(A)*, pp. 1160–1169, 1985.

[7] D. Heeger, “Model for the extraction of image flow,” *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 4, no. 8, pp. 1455–1471, Aug. 1987.

[8] L. Quam and M. J. Black. The yosemite sequence. [Online]. Available: http://www.cs.brown.edu/people/black/images.html

[9] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *IEEE Trans. Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[10] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” to appear in *IEEE Transactions on Image Processing*, Feb. 2006.

[11] E. P. Simoncelli and B. A. Olshausen, “Natural image statistics and neural representation,” *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.