Multi-scale and Scalable Video Quality Assessment
Digest of Technical Papers

Kalpana Seshadrinathan, *Student Member, IEEE*, and Alan C. Bovik, *Fellow, IEEE*

Abstract-- **With the rapid proliferation of digital video applications, the question of video quality control becomes central. We present a novel multi-scale framework for video quality assessment that models motion in video sequences and is capable of capturing spatio-temporal artifacts in digital video. Performance evaluation of the proposed metric on the VQEG database shows that the system is competitive with and even performs better than existing methods.**

I. INTRODUCTION

Digital video has pervaded the lives of people due to the popularity of applications such as Internet Video, Interactive Video on Demand (VoD), Video Telepresence, Video Phones, PDAs and other Wireless Video devices, Video Surveillance, HDTV, Digital Cinema etc. Unfortunately, each stage of processing that a video sequence goes through before reaching the end-user in any of these applications changes the *quality* of the video. Thus, algorithms that can automatically assess the quality of a video sequence are essential to monitor and control the quality of videos that are being distributed and communicated globally. In most of the applications mentioned above, the end users of the video sequence are human observers, who can instantaneously judge the quality of the video using their visual system. The goal of video Quality Assessment (QA) research is to predict the visual quality of a video signal, as assessed by a human observer. Most of the research on video QA has focused on quantifying the *fidelity* of a given video sequence, with respect to the original "perfect'" video sequence before any processing occurred, and this is known as Full Reference QA. We focus on full reference video QA in this paper.

In the literature, video quality assessment has always been addressed using simple modifications of models developed for still image QA [1,2,3,4]. The main reason for this has been the fact that motion processing in the Human Visual System (HVS) is not as well understood as the initial processing stages in the visual pathway that play a key role in human perception of static images. However, precisely due to the fact that motion processing occurs in the HVS, different factors come into play in QA of moving images, that are not addressed sufficiently by video QA systems that are based on still image QA systems. We have described the importance of modeling motion and temporal artifacts in video sequences and demonstrated gains using such an approach [5,6].

In this paper, we will develop a multi-scale framework for video quality assessment, that improves upon the single-scale algorithms that we have previously developed for video QA [5,6].

II. MULTI-SCALE VIDEO QA

Most video QA systems in the literature employ a scale-space decomposition, usually separable, of the image/video signal to mimic similar processing in early stages of visual processing. Example spatial decompositions include the Cortex transform and the steerable pyramid [7,8]. Temporal frequency decomposition is performed using either a single or two channel model, that attempts to model temporal processing by neurons in the visual cortex [1,3]. However, such simple temporal processing is insufficient to characterize the response of neurons in the Medial Temporal Area (Area MT) of the visual cortex that is well known to play a critical role in movement perception in the HVS.

The QA models in [5,6] perform a decomposition of both the reference and test video sequences into *spatio-temporal* bandpass channels in the frequency domain that differ significantly from others used in video QA. This decomposition achieves two goals: optical flow estimates, that describe the motion of each pixel in the video sequence as a two-dimensional vector, are derived using the outputs of these bandpass channels. Secondly, the video quality is computed between these bandpass filtered outputs in the frequency domain, as opposed to the pixel domain. A family of Gabor filters at a *single scale* was used in our implementation of both quality metrics and an iso-surface contour of the Gabor filterbank in the spatio-temporal frequency domain is illustrated in Figure 1. The spatio-temporal decomposition that we use is selective for the velocities of visual stimuli and has been successfully used for optical flow estimation on video sequences [9]. Further, filters such as ones we have used have also been proposed as physiologically plausible models for the tuning of neurons in Area MT [10]. To the best of our knowledge, the models in [5,6] are the first to use decompositions that are velocity-selective in a video quality assessment framework. However, all the filters used in [5,6] were at a single scale, which results in several disadvantages that we outline below. In this paper, we attempt to overcome these limitations by developing a multi-scale set of velocity-selective filters.

The first and most significant drawback of a single scale filterbank is the inability to detect motion that causes the spectrum of the video to lie outside the bandpass support of the filters. We use the Fleet and Jepson optical flow estimation algorithm and all reported implementations of this algorithm deploy a single scale of filters [9]. Such filters hence fail to

Thanks to the National Science Foundation for funding.

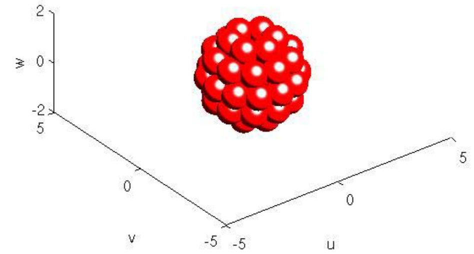

Figure 1: Iso-surface contours of the Gabor family in the spatio-temporal frequency domain

compute optical flow in fast moving regions of videos, since fast moving regions need to be detected at lower spatial frequencies to avoid the effects of temporal aliasing [9]. This is a drawback for video QA, since videos commonly contain fast moving objects (sports, action movies, etc.). Challenges that a multi-scale framework will encounter include automatic selection of the scale at which motion is to be detected, and detecting and avoiding filter outputs that suffer from temporal aliasing.

In addition to motion estimation issues, single scale filterbanks do not span the entire frequency domain. This is also not desirable in a QA framework, since spectral components of *distortions* in the video, may well fall outside the passbands of the filterbank and will hence not be detected. This is all the more important in applications such as compression, where quantization causes loss of high frequency information, that cannot be detected by the filters we use in [5,6]. Additionally, a multi-scale framework can be used to model a number of perceptually-relevant effects, such as the reduced visibility of spatial detail in fast moving regions, the high visibility of flickering artifacts, and so on. To account for these, we propose to compute quality indices using only the filters at the scales at which motion is detected. We propose to demonstrate the performance of our multi-scale framework for QA using the Video Quality Expert's Group (VQEG) database [13].

III. Scalable Image and Video QA

A multi-scale framework can be used in *scalable* image and video QA and to incorporate adjustments for viewing distance [11]. The spatial scale (spatial resolution) and temporal scale (frame rate) of a video stream are often altered by, e.g., display or transcoding requirements. It is therefore of interest to perform QA on images and videos that have been scaled relative to the reference. Example applications include scalable streaming video over the Internet, video display on small mobile devices, in-flight entertainment screens, High Definition video displayed on Standard Definition monitors, etc. Our approach to scalable QA will begin with still images, since that problem remains unaddressed and is a suitable precursor to scalable video QA. Since test and reference images are both resolution decomposed, the resolution scales can be made to match using filters at different scales. If the test image is *1/s* the size of the reference (for simplicity, assume the same reduction in scale along horizontal and vertical axes), then the subband filters operating on the test image will likewise be scaled by a factor *1/s* relative to those used on the reference. Subsampling the reference by a factor *s* after resolution decomposition will then produce a scale matched reference, which can be used as the reference signal in the QA algorithm. Such approaches have been used previously in scale-matched object recognition [11].

We will describe techniques for scalable QA of still images and demonstrate the performance of our proposed technique. We will also briefly discuss ways to extend these scalable techniques to video QA.

References

[1] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 2668, no. 1. San Jose, CA, USA: SPIE, Mar. 1996, pp. 450–461.

[2] 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.

[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] K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, January 25-26 2007.

[6] K. Seshadrinathan and A. Bovik, "A structural similarity metric for video based on motion models," in *2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, pp. I–869–72.

[7] A. B. Watson, "The cortex transform: rapid computation of simulated neural images," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 311–327, 1987.

[8] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 587–607, 1992.

[9] 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.

[10] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area mt." *Vision Res*, vol. 38, no. 5, pp. 743–761, Mar 1998.

[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2003 Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, pp. 1398–1402, 2003.

[12] D. Ballard and L. Wixson, "Object recognition using steerable filters at multiple scales," in *Proceedings of IEEE Workshop on Qualitative Vision*, pp. 2–10, 1993.

[13] (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/index.php