DCT-based Video Quality Evaluation

---Final Project for EE392J

Feng Xiao

Winter 2000

Abstract

Lossy compression methods make the widespread distribution of digital video possible at the cost of introduction of artifacts. To control the introduction of potentially visible artifacts, a proper automated video quality evaluation method is needed. Based on its simplicity, pixel-based Root Mean Square Error(RMSE) or its derivatives(PSNR or SNR) is dominant metric in practice. However, RMSE doesn't take into account of the spatial-temporal property of human's visual perception that is the reason why it fails under many circumstances. In this project, I have developed a modified DCT-based video quality metric (VQM) based on Watson's proposal, which exploits the property of visual perception. This metric uses the existing DCT coefficients, so it only incurs slightly more computation overhead.

First, I show how RMSE fails in some common situations and VQM can account for these. Second, I try to find how these widely used compression options (quantization parameter, quantization matrix, spatial scalability, frame dropping) affect the video quality.

1. Introduction

An uncompressed sequence of digital video can occupy a vast amount of storage space and bandwidth. For example, the typical CIF format for MPEG-1 is 360 pixels width and 288 pixels height. Assuming it is in color with 8-bit precision, each picture occupies about 311 Kbytes. If the frame rate is 24 pictures/second, then the raw data is about 60Mbit/s. Most recent commercial bandwidth is far below this bit-rate.(33.6 for modem, n*64 for ISDN, 1M~ for ATM). So high compression ratio is critical for widespread video distribution.

Compression methods can be divided into two categories: lossless and lossy compression. While lossless methods alone (Huffman coding for example) can only achieve a 3:1 ratio at best, lossy methods can achieve a much high compression ratio (100 or higher). As a result, lossy methods always introduce artifacts. To control the visual artifacts associated with lossy compression, a proper video quality evaluation metric is necessary. Unfortunately, all these recent digital video compression standards(MPEG-1,2,4, H.261, H.263 and so on) use RMSE-based metric(SNR or PSNR), which is too simple to be a sound means for measuring visual quality.

Human spatial-temporal contrast sensitivity function (Figure 1) is the base for all sound video metrics. From figure 1, we can see that approximately eyes' sensitivity to spatial-temporal pattern decreases with high spatial and temporal frequency. Based on the different sensitivity, we can represent high spatial or temporal information with less

precision while human's eyes are not sensitive to the loss of this information. DCT quantization exploits this property directly.

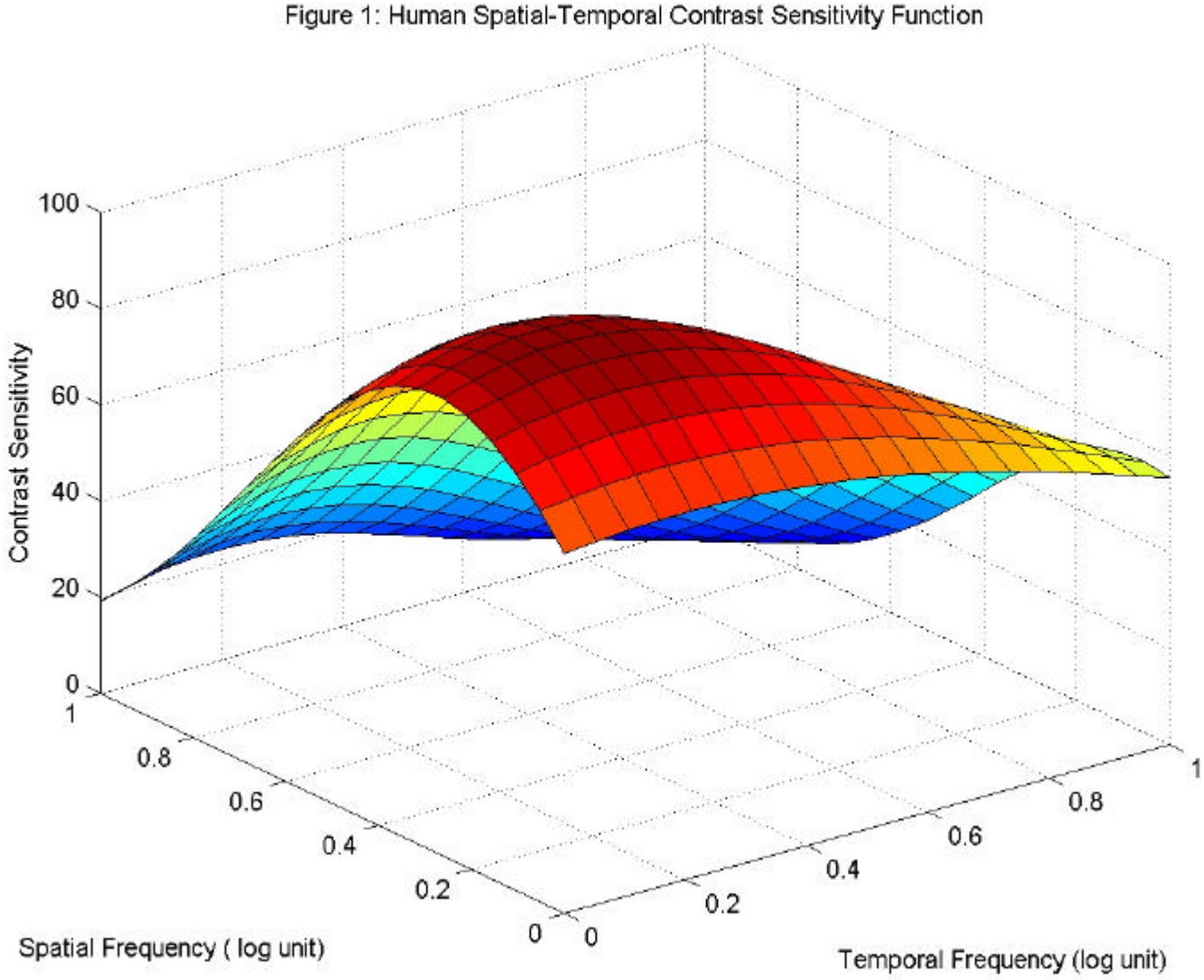


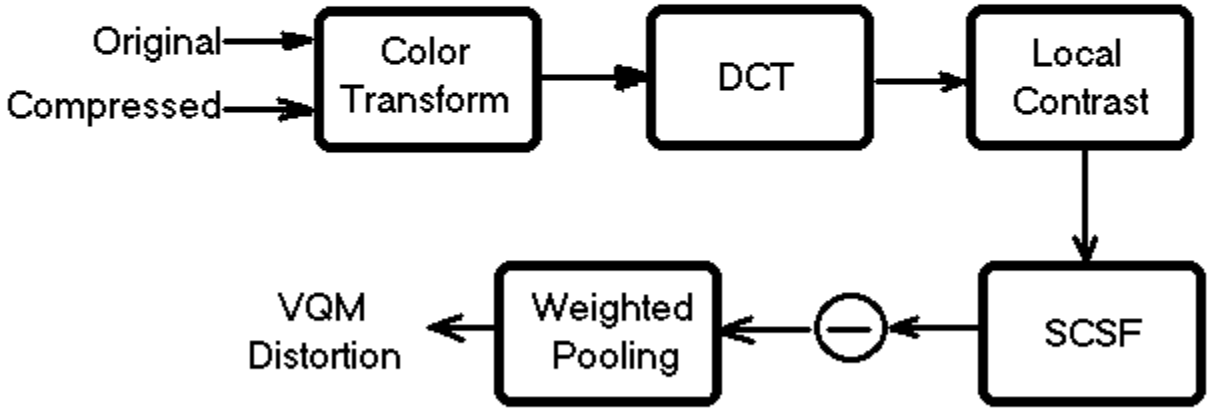
Figure 1. Human Spatial-Temporal Contrast Sensitivity Function

Recently a number of video quality metrics [ref.1,2,3,5,6,8,9,10,11,12] have been proposed as alternatives for RMSE, but most of them may not be able to satisfy two constrains at the same time: closely enough upon human perception and computation simple. Watson's DCT-based metric (DVQ) [ref. 9] seems to be the most promising candidate because its use the DCT coefficients make it efficient while it is much closer to human perception than RMSE. Based on his DVQ model, I develop my own DCT-based metric (VQM).

2. VQM Overview

Figure 2 is an overview of the flowchart of VQM.

Figure 2. Overview of QVM



2.1 Color transform. Both MPEG and H.263 use the YUV color space, so it can use the raw data directly.

2.2 DCT transform. Approximately said, this step separates incoming images into different spatial frequency components.

2.3 Converts each DCT coefficients to local contrast (LC) using following equation:

$$LC(i,j) = DCT(i,j) * Power(DC/1024, 0.65) / DC$$

DC is the DC component of each block. For 8-bit image, 1024 is mean DCT value. 0.65 is the best parameter for fitting psychophysics data. After this step, most values lie between [-1,1].

The above three steps are identical to Watson's DVQ model.

2.4 Converts LC to just-noticeable differences (jnds).
This step is different from Watson's model. Instead of applying temporal filtering and human spatial contrast sensitivity function (SCSF) separately, I choose to apply one SCSF matrix for static frames and one matrix for dynamic frames in one step. This will further reduces the computation and memory load. The DCT coefficients are converted to just-noticeable differences by multiplying each DCT coefficient by its corresponding entry in the SCSF matrix. For static SCSF matrix, I choose the MPEG default

quantization matrix (See appendix, quantization matrix can be seem as the contrast threshold matrix, which is the inverse of contrast sensitivity matrix). as its inverse since this matrix is based on extensive psychophysics research. For dynamic matrix, I choose to raise each entry in static SCSF matrix to a power to account for the temporal property of SCSF. The power is decided by the frame-rate of video sequences.

2.5 Weighted pooling of mean and maximum distortion.
The two sequences are subtracted first. At this step VQM also differs from DVQ by incorporating contrast masking into a simple maximum operation and then weights it with the pooling mean distortion. This reflects the facts that a large distortion in one region will suppress our sensitivity to other small distortion, for this kind of situation, weighted maximum distortion into pooled distortion is much better than pooled distortion alone.

Mean_Dist= 1000*mean(mean(abs(diff))).
Max_dist= 1000*maximum(maximum(abs(diff)))
VQM= (Mean_dist+0.005*Max_dist).

Maximum distortion weight parameter 0.005 is chosed based on several primitive psychophysics experiments. Parameter 1000 is the standardization ratio.

3. VQM vs. RMSE

RMSE is based on the assumption that human observer is sensitive to the summed squared deviations between reference and test sequences, and is insensitive to other aspects like spatial and temporal frequency or color of the deviations. This extremely simple assumption restricts its usage. It fails for some common circumstances.

In the first experiment, I build a series of images by adding different spatial frequency noise and compute their VQM and RMSE distortion regarding to original image. Figure 3 shows several of these noise images and the original image. From observer's view, these images have very different levels of distortion while the RMSE distortion is the same. VQM distortion is proportional to our impression.

As block data is the atom of DCT-base video processing, frequent loss or degrade of it causes many block-based distortions. We can see this kind of distortion in MPEG or H.263 applications frequently. A good video quality metric should be able to account for it. In the second experiment, I add different levels of noise to one block and compute the normal pooling DCT distortion and the revised pooling distortion with weighted maximum distortion. It shows that the latter is more sensitive to this kind of distortion, thus is more consistent with our subjective distortion (Figure 4).

In both experiments, the threshold for detecting distortion is approximate 1 unit. This number can serve as an approximate index for evaluating distortion.

4. VQM with scalability

Not all video applications have a single, well-defined end user. Video Services such as ATM, and HDTV with TV backward compatibility need to deliver more than one resolution and quality. Internet-based video delivery needs to be adapted to network traffic by dropping frames or so. This kind of applications calls for scalability. How to achieve the best video quality by allocating bit-rate to different parts is critical for video scalability. In the following experiments, I try to evaluate video quality distortion for several commonly employed compression options:

a. Quantization matrix vs. single quantization parameter
b. Different quantization levels
c. High-spatial resolution vs. low-spatial resolution
d. Dropping frames

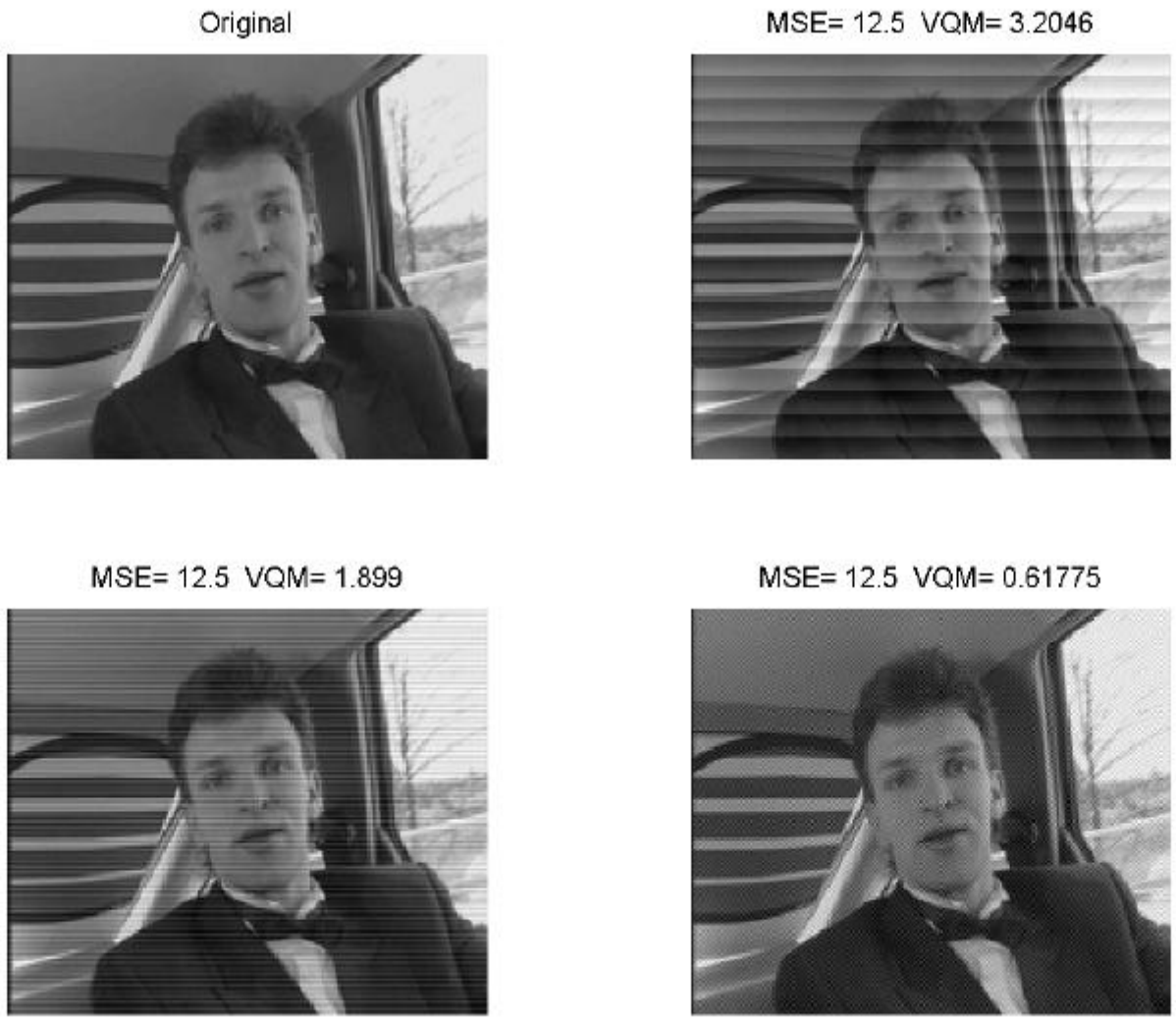


Figure 3. VQM vs. RMSE

Figure 4: Weighted maximum distortion vs. mean only

4.1 Quantization matrix vs. single quantization parameter
As we know, MPEG-1 chooses a fixed 64-item quantization matrix (the one that I choose for static frame SCSF) for intra-code and a single quantization parameter for inter-code. MPEG-2 chooses the same matrix as the default matrix but extends to download user-supplied matrix at fly. H.263 targets a much lower bit-rate (~30 kbits/s) than MPEG (~1.5 Mbits/s or higher), so it uses a single quantization parameter for both intra-code and inter-code. It is worthwhile to know how much gain we can get by adding this overhead.

In figure 5, it shows that VQM is approximate the same when single quantization parameter equals 24, which is **2/3** of the mean of the 64-item quantization matrix. This may indicate that an optimized matrix performances only slightly better than a flat matrix

(degrades to a single parameter), the gain is definitely less than 1.5(3/2) even if we don't count the overhead incurring by download matrix. This may be the reason why people prefer a single parameter to a matrix in H.263. Since I don't have the MPEG encoder, I cannot show the exact gain for choosing a quantization matrix over a single parameter (flat matrix).

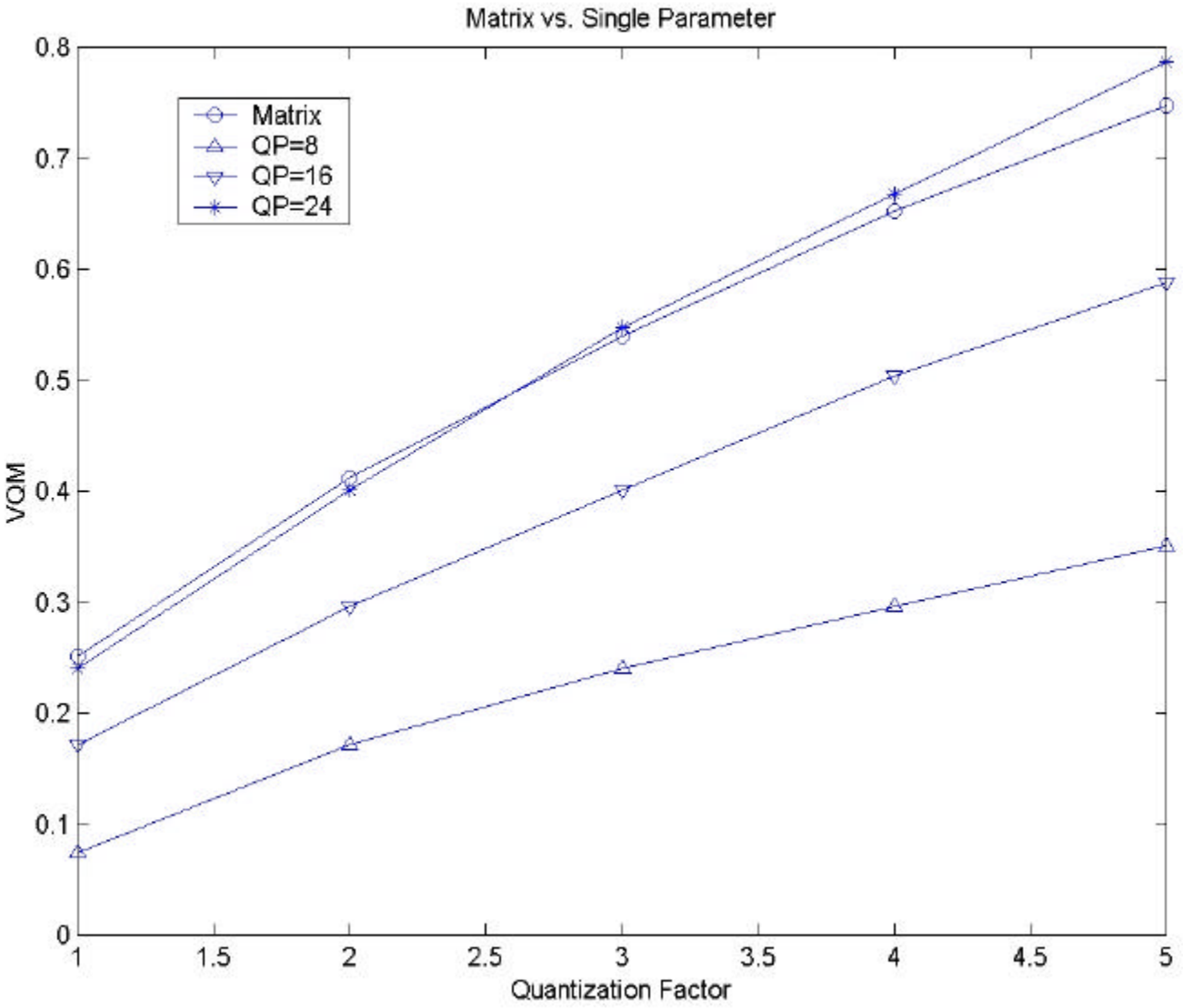


Figure 5: Quantization Matrix vs. Single Quantization Parameter

4.2 Different quantization levels

Using different quantization parameter is the simplest way to control bit-rate. It will be interesting to investigate the VQM distortion associated with different quantization parameters.

From figure 5, we can see the relationship between VQM distortion and quantization factor is approximately linear.

4.3 High-spatial resolution vs. low-spatial resolution

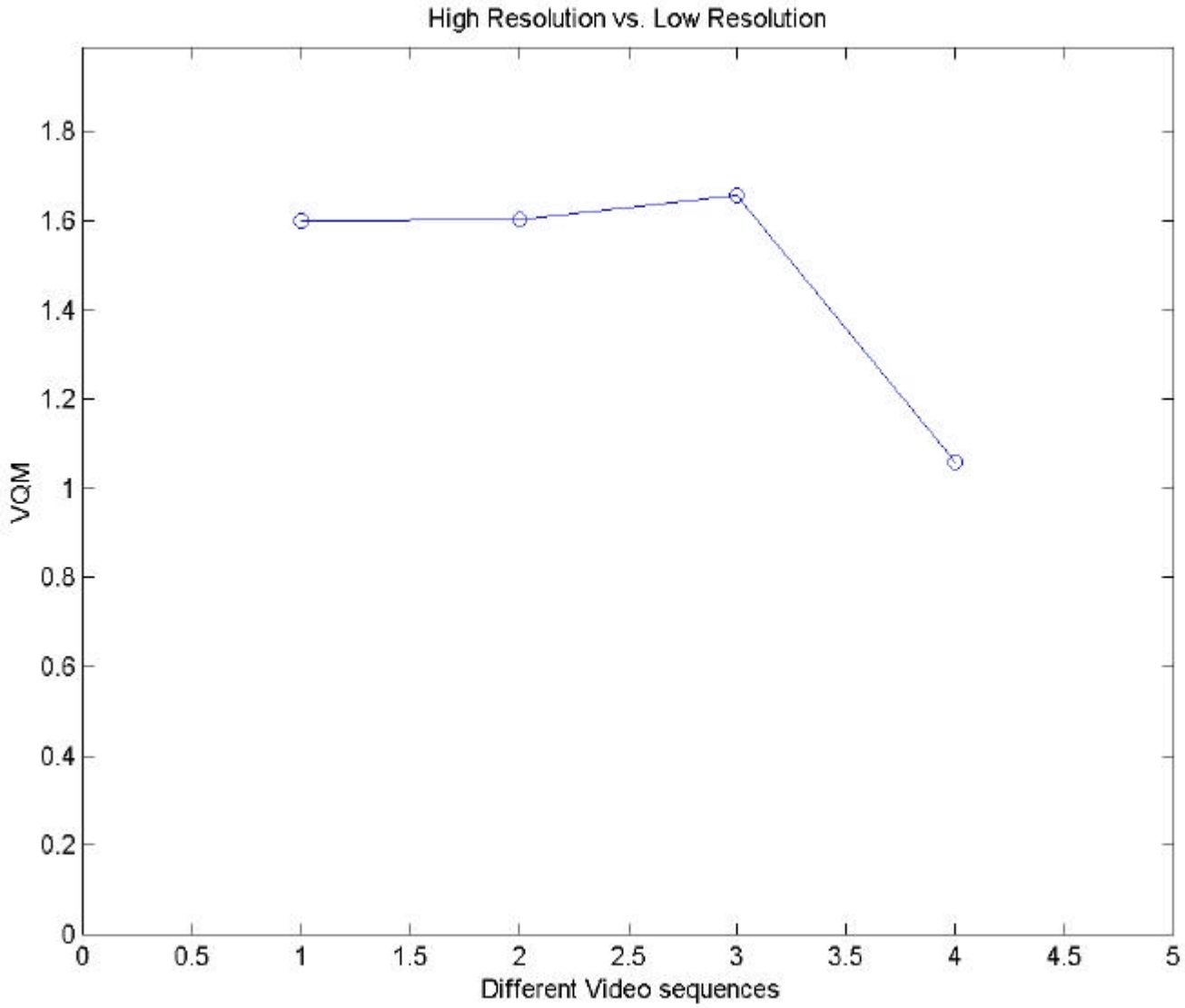


Figure 6: high-resolution image and low-resolution image

From left to right (carphone, foreman, grandma, claire)

Sometimes, we may want to trade bit-rate with high-spatial resolution. In this experiment (figure 6) I compute the distortion between high-resolution (144 by 176) and low-resolution clips(72 * 88) for four kinds of video sequences. The distortion is barely above the distortion index (varies slightly with sequences).

4.4 Dropping frames.

In H.263, dropping frames is also widely used. In this experiment, distortion is computed for original sequences and dropping sequences. The dropped frame number between two

consecutive frames varies from 1 to 9 (if the original frame rate is 30 frames/sec, then resulting frame rate will be 15 frames/sec to 3 frames/sec respectively).

It is quite clear that distortion caused by dropping is context sensitive. It varies a lot for different video sequences. Based on the visible distortion threshold (approximately 1 unit), we cannot drop any frames for foreman and carphone sequences, while we can drop one frame for grandma sequence and up to 9 frames for claire sequence without noticeable distortion.

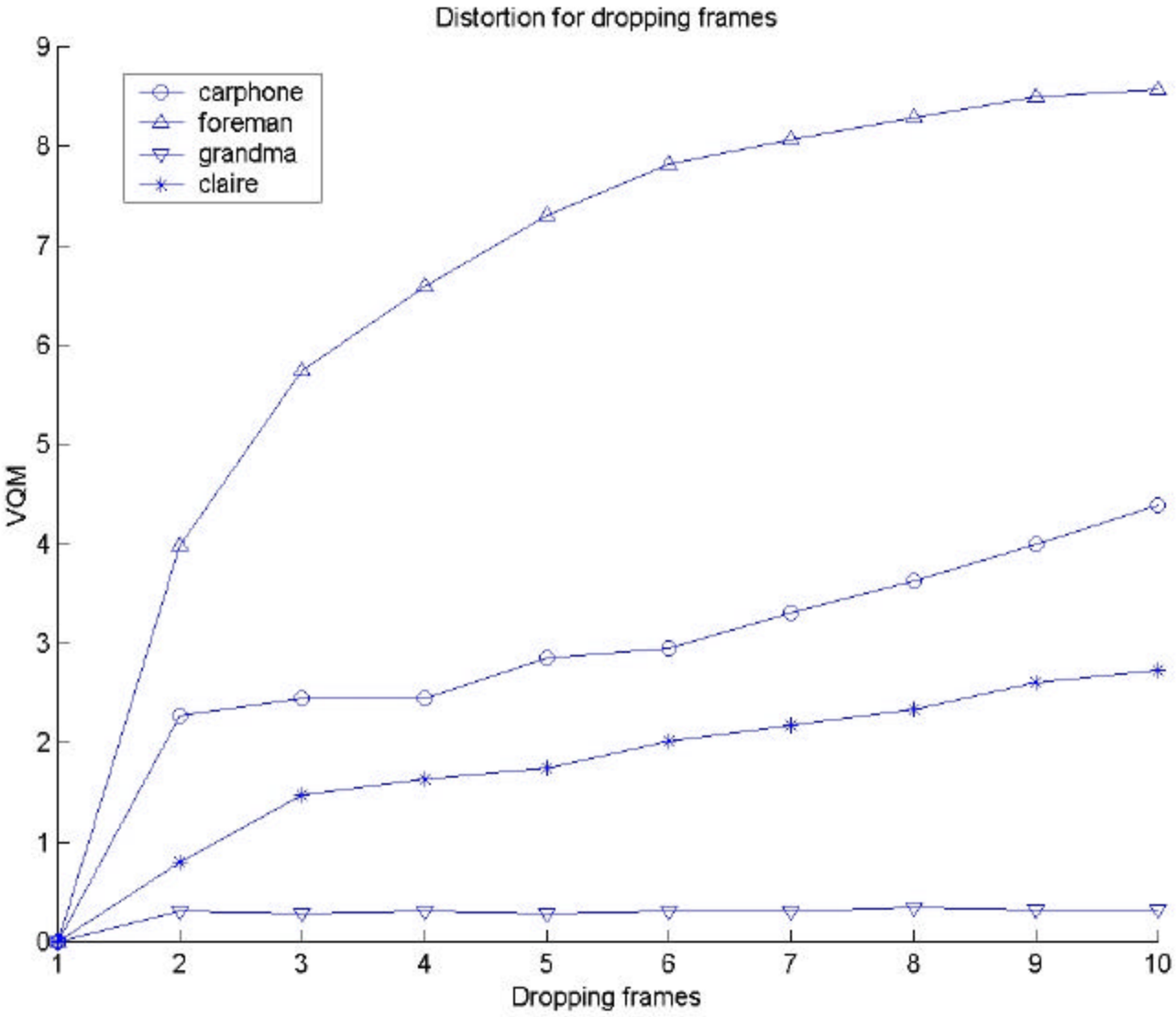


Figure 7: Distortion when dropping frames

5. Summary

I have describe the simple DCT-based video quality evaluation metric (VQM) which is based on a simplified human spatial-temporal contrast sensitivity model. Comparing to

RMSE-based metrics, it performances much better in these situations when RMSE fails. The light computation and memory load make it even more attractive for wide applications. Also, I evaluate how these frequently used compression technology (quantization matrix, spatial scalability, temporal scalability) affect video distortion. The model and the results are very primitive, it needs more works in the future.

Acknowledgment: I wish to thank EE392J's instructors Dr. John Apostolopoulos and Dr. Susie Wee for providing me H.263 source code and guidance for this project. Thank TA Hareesh Kesavan for the discussion of H.263 and technical help. Thank my teammates Suiqiang Deng and Bao-jun Jiang.

6. References

1. T. Hamada, S. Miyaji and S. Matsumoto, "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception", Society of Motion Picture and Television Engineers, 179-192 (1997).
2. C.v.d.B Lambrecht. "Color moving pictures quality metric", International Conference on Image Processing,I, 885-888 (1996).
3. J.Lubin, "A Human Vision System Model for Objective Picture Quality Measurements",International Broadcasters' Convention, Conference Publication of the International Broadcasters' Convention,498-503(1997).
4. J.L.Mitchell,W.B.Pennebaker,C.E.Fogg and D.J. LeGall. "MPEG video compression standard"International Thomson Publishing, 1997.
5. K.T.Tan, M.Ghanbari and D.E. Pearson, "A video distortion meter", Picture Coding Symposium, 119-122 (1997).
6. X.Tong, D.Heeger and C.v.d.B. Lambrecht. "Video Quality Evaluation Using ST-CIELAB", Human Vision, Visual Processing and Digital Display, SPIE Proceedings, 3644, 185-196 (1999).
7. Brian.A.Wandell. "Foundation of Vision", Sinauer Associates, Inc, Sunderland, Massachusetts, 1995.
8. A.B. Watson, "Image data compression having minimum perceptual error", US Patent 5,629,780. (1997)
9. A.B. Watson, "Toward a perceptual video quality metric", Human Vision, Visual Processing, and Digital Display VIII, 3299, 139-147 (1998).
10. A.B. Watson, J.Hu, J.F.McGowan III and J.B. Mulligan. " Design and performance of a digital video quality metrix". Human Vision and Electronic Imaing III, SPIE Proceedings, San Jose, (1998).
11. A.A.Webster,C.T.Jones,M.H. Pinson,S.D.Voran and S. Wolf, " An objective video quality assessment system based on human perception", Human Vision, Visual Processing, and Digital Display IV, SPIE Proceedings, 1913, 15-26 (1993).
12. S.Wolf,M.H. Pinson, A.A.Webster,G.W. Cermak and E.P. Tweedy, "Objective and subjective measures of MPEG video quality", Society of Motion Picture and Television Engineers, 160-178 (1997).

Appendix

MPEG default quantization matrix table:

```
[8 16 19 22 26 27 29 34;
16 16 22 24 27 29 34 37;
19 22 26 27 29 34 34 38;
22 22 26 27 29 34 37 40;
22 26 27 29 32 35 40 48;
26 27 29 32 35 40 48 58;
26 27 29 34 38 46 56 69;
27 29 35 38 46 56 69 83];
```