VIDEO QUALITY MEASURES BASED ON THE STANDARD SPATIAL OBSERVER

A.B. Watson*

NASA Ames Research Center
Vision Group
MS-262 Moffett Field, CA, 94035, USA
abwatson@mail.arc.nasa.gov

J. Malo†

Universitat de València
Dept. d'Òptica
Dr. Moliner 50. Burjassot 46100, SPAIN
jesus.malo@uv.es

ABSTRACT

Video quality metrics are intended to replace human evaluation with evaluation by machine. To accurately simulate human judgement, they must include some aspects of the human visual system.

In this paper we present a class of low-complexity video quality metrics based on the Standard Spatial Observer (SSO). In these metrics, the basic SSO model is improved with several additional features from the current human vision models.

To evaluate the metrics, we make use of the data set recently produced by the Video Quality Experts Group (VQEG), which consists of subjective ratings of 160 samples of digital video covering a wide range of quality. For each metric we examine the correlation between its predictions and the subjective ratings.

The results show that SSO-based models with local masking obtain the same degree of accuracy as the best metric considered by VQEG (P5), and significantly better correlations than the other VQEG models. The results suggest that local masking is a key feature to improve the correlation of the basic SSO model.

1. INTRODUCTION

Video coding techniques are designed to optimize the rate-distortion performance of the encoder for a better use of the available bandwidth in the communication channel [1]. In many applications the final user of the visual information is a human observer. Therefore the distortion measure used in the rate-distortion design has to be meaningful in subjective terms.

It is well-known that simple energy based metrics such as the Mean Square Error (MSE) are not suitable to describe the subjective degradation perceived by a viewer. To counter the limitations of MSE, designers often include some aspects of models of early human vision in order to obtain better correlations between the predictions of the metric and the opinion of human observers [2, 3, 4]. However, these metrics are often computationally expensive, and the relative value of each element of the model remains obscure.

The aim of this paper is to present a class of low complexity distortion metrics based in the Standard Spatial Observer (SSO) [5, 6]. Here the basic SSO model is improved using several additional features from current human vision models (temporal frequency sensitivity, temporal smoothing and summation, masking, and a simple model that accounts for the subjective effect of field duplication).

The different proposed metrics are evaluated according to the VQEG recommendations [7].

This analysis has both theoretical and applied interest. From the theoretical point of view it is interesting to see which aspect of the models is more important to give an adequate description of the observer behaviour in a complex task involving natural images. This is particularly important because in the basic literature each feature is studied separately using lab stimuli. Therefore, it is difficult to assess the relevance of each feature in natural conditions. From the applied point of view, it is important to see whether the addition of features can improve the basic performance of the SSO, and also whether simple models such as the SSO are able to perform as well as other (more complex) VQEG models. Besides, the ranking of the features of the model according to their impact in the correlation will give the necessary information to design the best behaved metric for a given computational cost.

2. DISTORTION MEASURES WITH NESTED ARCHITECTURE

Starting from a simple energy difference, such as the MSE, we have included sequential refinements of this measure using the basic ingredients of the current mod-

*ABW was supported by NASA Grant 00-HEDS-01-055.
†JM was supported by MEC-Fulbright Grant FU29167406.

els of early human vision [8]:

- Standard Spatial Observer [5, 6]:
  - Operates on contrasts rather than luminances.
  - Spatial contrast sensitivity function.
  - Non quadratic norm for spatial summation.
- Temporal contrast sensitivity function [9].
- Second stage temporal filter prior to temporal summation.
- Non quadratic norm for temporal summation [9].
- Local masking [8].
- Field replication model.

This modular (nested) structure allows a balanced trade-off between accuracy and complexity.

3. IMPLEMENTATION OF SSO-BASED MODELS

The Standard Spatial Observer (SSO) is a very simple algorithm that evaluates the perceptual distance between a pair of 2D contrast patterns. SSO-based video metrics use this model (and some extensions of it) to assess the visibility of the difference between the original sequence and the distorted sequence,

$$d(x,t) = s_o(x,t) - s_d(x,t) \tag{1}$$

The subjective distortion, d, (i.e. the visibility of the difference contrast pattern) in the plain SSO model is computed by a quadratic temporal summation of the differences, $d(t)$, in each frame:

$$d = \left(\sum_t d(t)^{\beta_t} \right)^{1/\beta_t} \tag{2}$$

with $\beta_t = 2$. The difference in each frame is given by the β_x-norm of the visible frame difference, $d(x,t)_{sso}$:

$$d(t) = \left(\sum_x d(x,t)_{sso}^{\beta_x} \right)^{1/\beta_x} \tag{3}$$

Here the Minkowski summation exponent is $\beta_x = 2.9$. In this model, the visible frame difference is a filtered version of the actual frame difference:

$$d(x,t)_{sso} = FT_x^{-1} \left(CSF_x \cdot FT_x \left(d(x,t) \right) \right) \tag{4}$$

where FT_x stands for spatial Fourier Transform and CSF_x is the spatial Contrast Sensitivity Function including the oblique effect.

Now we list the set of features we added to the basic SSO measure. They are denoted individually by single lower case letters (t, p, m, and h), but of course we also used them jointly.

3.1. Temporal Pre-Filter Before the Spatial Summation (SSO+t)

This extension includes a temporal frequency response, CSF_t. In this case the visible frame difference is:

$$d(x,t)_{sso+t} = FT_t^{-1} \left(CSF_t \cdot FT_t \left(d(x,t)_{sso} \right) \right) \tag{5}$$

We tried three values for the time constant (or cut frequency) of the CSF_t filter: the standard one (the one to reproduce the Robson CSF_t [9]), one bigger (by a factor 2) and one smaller (by a factor 2). We found that (in the initial 25 sequences set) the best result was obtained with the Robson-like CSF_t.

3.2. Temporal Post-Filter After the Spatial Summation (SSO+p)

In this modification we jointly optimized the temporal summation, β_t in eq. 2, and a low-pass temporal filter, F_t, applied after the spatial summation to remove high frequency oscillations in $d(t)$. In this case, we use $d(t)_{sso+p}$ in eq. 2, where:

$$d(t)_{sso+p} = FT_t^{-1} \left(F_t \cdot FT_t \left(d(t) \right) \right) \tag{6}$$

3.3. Local Masking (SSO+m)

In this extension the visible difference frame is not $d(x,t)_{sso}$, but a masked version of it:

$$d(x,t)_{sso+m} = \frac{d(x,t)_{sso}}{\sqrt{1 + \left(\frac{h(x,t) * s_o(x,t)}{c} \right)^2}} \tag{7}$$

In this masking model each spatio-temporal difference is attenuated by the local energy of the original pattern. We crudely optimized the size of the neighbourhood (the width of the Gaussian kernel, h) and the strength of the masking effect using the constant, c, for a 90 sequences set. We found that:

- The best width of the kernel is around twice the size of the CSF_x impulse response.
- The optimum c implies a well-behaved non-linearity (similar to the Naka-Rushton curve [8]).

We also tried global masking (i.e. considering the energy of the whole frame and optimizing c), but it performed poorly, so we decided to exclude those results from this discussion (+m always means local masking).

3.4. Field Replication Adjustment (SSO+h)

Some video processing schemes (e.g. H.263) duplicate individual fields of the two-field video frame. This introduces a spatial misalignment between the pixels of

the original and the distorted sequence. In distortion measures based on pixel-by-pixel comparisons, this effect introduces a dramatic overestimation of the distortion. This overestimation is specially severe for sequences with rapidly moving objects.

We have developed an algorithm for detecting and compensating for this over-estimation. Since our focus in this paper is on vision models, and this is a specialized non-visual algorithm, we only describe it briefly. The essence of the algorithm is to detect field duplication, estimate the over-estimation of error, and correct it. The overestimation can be deduced from the amount of motion in the scene.

Note that when this procedure is used, it is applied after all the processes are done, so sometimes it may reduce the efficiency of previously optimized processes.

4. METHODS

The proposed algorithms were optimized and tested on the 30Hz VQEG set to maximize the correlation between the predictions of the models, d, and the experimental Differential Mean Opinion Score, $DMOS$. The 30Hz VQEG set consist of 160 stimuli (10 sequences times 16 possible distortions). In order to decouple optimization and evaluation, we used different subsets and different correlation measures in both cases.

4.1. Optimization of the Parameters

Some of the extensions have parameters. The final values were chosen using a non-exaustive optimization. These optimizations were done using a discrete parameter space and using a restricted subset of (25-90) sequences. They were optimized to maximize C, a sum of parametric and non-parametric correlation measures:

$$C = P(\chi^2 > \epsilon_v^2) + \rho_s \tag{8}$$

where ρ_s is the (non-parametric) Spearman correlation, and $P(\chi^2 > \epsilon_v^2)$ is the χ^2 parameter of the fit of $DMOS = f(d)$ where f is the sigmoid II recommended in VQEG [7]. In this case, ϵ_v^2 is just the sum of squared errors taking into account the variances.

The individual extensions were optimized one at a time. In the measures obtained using combinations of the extensions, the parameters were not re-optimized.

4.2. Comparison between Models

The comparison between the SSO based models and the P0-P9 distortion metrics proposed in the VQEG project was done according to the root mean square (RMS) error recommended by VQEG. For a given model, j, the RMS error, ϵ_j, is:

$$\epsilon_j = \left(\frac{1}{160}\sum_{i=1}^{160}\epsilon_{ij}^2\right)^{1/2} = \left(\frac{1}{160}\sum_{i=1}^{160}(DMOS_i - d_{ij})^2\right)^{1/2} \tag{9}$$

The ϵ_j numbers may give a ranking of the compared models as shown in section 5. However, the VQEG recommendations do not make clear how to decide when the differences in ϵ_j are significant.

In order to solve this we used the Fisher-Snedecor test on each pair of deviations, ϵ_j and ϵ_k, to decide if they are equal or different. This test is based on the fact that given two χ^2-like independent variables (such as ϵ_j^2 and ϵ_k^2) the quotient,

$$Q_{jk} = \frac{\epsilon_k^2}{\epsilon_j^2} \tag{10}$$

is distributed according to a Fisher-Snedecor PDF. Once this is known, one can compute the probability $P(Q_{jk} > 1) = P(\epsilon_k > \epsilon_j)$ In our case, the independence was determined using the Spearman correlation ($\rho_s < 0.5$) between ϵ_{ij}^2 and ϵ_{ik}^2.

The independence condition does not allow a strict comparison between very similar models (which of course are not independent). However, it may be applied to independent-enough models (such as any SSO based model and any P0-P9 VQEG models). By doing this, one can get a sense of the sort of difference in ϵ_j which is significant.

5. RESULTS AND DISCUSSION

Table 1 shows the RMS error for the 30 distortion metrics considered in this study, ϵ_j with $j = 1, \cdots, 30$. A useful reference is the P0 model in VQEG (the standard PSNR).

The features described in section 3 have also been applied to the MSE difference because, in principle, it is similar to the SSO measure: MSE is just the SSO with no CSF filter and quadratic spatial summation. However, note that the MSE based results are poorer than SSO. This suggests that the CSF filtering and the spatial summation in SSO have a substantial effect.

Also note that the interactions between the different individual extensions of the SSO model are not trivial.

Figure 1 shows the $P(\epsilon_k = \epsilon_{j_0})$ for different models j_0. Values above the straight line mean that these ϵ_k are not significantly different from ϵ_{j_0}. The first (left hand) solid line means the probability of each ϵ_k (or model) to be worse than the best model (the one with lowest ϵ). When this line drops below the significance

	Model	ϵ_j
1	SSO+m+h	0.066
2	SSO+m+p+h	0.066
3	P5	0.068
4	SSO+t+m+p+h	0.068
5	SSO+t+m+h	0.068
6	SSO+m+p	0.069
7	SSO+t+m+p	0.069
8	SSO+t+m	0.071
9	SSO+m	0.072
10	SSO+p+h	0.074
11	MSE+m	0.076
12	P1	0.076
13	SSO+p	0.076
14	SSO+h	0.076
15	SSO	0.079
16	SSO+t+p+h	0.081
17	P2	0.081
18	SSO+t+p	0.082
19	SSO+t+h	0.082
20	SSO+t	0.084
21	MSE	0.085
22	P0 (PSNR)	0.086
23	P9	0.087
24	P3	0.092
25	P7	0.093
26	P8	0.096
27	MSE+m+p	0.098
28	P4	0.103
29	MSE+m+p+h	0.106
30	P6	0.154

Table 1. Model ranking for 160 VQEG 30Hz sequences.

value ($P = 0.1$) the model is worse than the best. This happens for the model in 10^{th} place, so the results of the first nine models are statistically equivalent. The same analysis can be done for the 10^{th} model (dashed line). In this case, the first model which is significantly worse is the 20^{th} model. This procedure can be applied in the same way for the other significantly different models ($j_0 = 20, 25, 29$). It allows us to give some illustrative boundaries between significantly different models (as shown in table 1).

6. CONCLUSION

The results show that SSO-based models with local masking attain the same degree of accuracy as the best VQEG model (P5) and significantly better correlations than many other VQEG models. The results also suggest that local masking is a key feature to improve the performance of the basic SSO model.

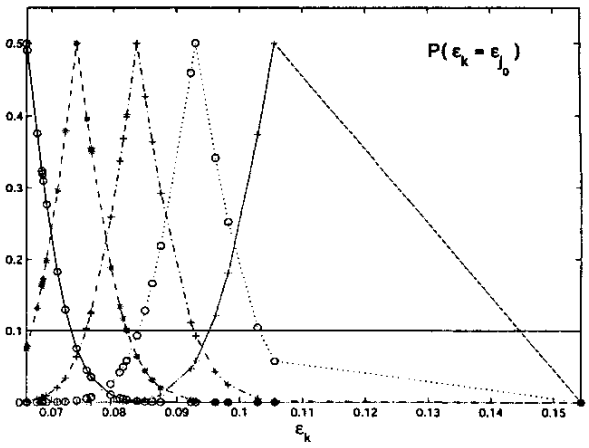

Fig. 1. $P(\epsilon_k = \epsilon_{j_0})$ with $j_0 = 1, 10, 20, 25, 29$.

7. REFERENCES

[1] J. Malo et al., "Perceptual feedback in multigrid motion estimation using an improved DCT quantization," IEEE Trans. Im. Proc., vol. 10, no. 10, pp. 1411–1427, 2001.

[2] A.J. Ahumada, "Computational image quality metrics: A review," in Intl. Symp. Dig. of Tech. Papers, 1993, vol. 24 of Proc. SID, pp. 305–308.

[3] J. Malo, A.M. Pons, and J.M. Artigas, "Subjective image fidelity metric based on bit allocation of the HVS in the DCT domain," Im. Vis. Comp., vol. 15, no. 7, pp. 535–548, 1997.

[4] A.B. Watson, J. Hu, and J.F. McGowan, "Digital video quality metric based on human vision," J. Electr. Imag., vol. 10, no. 1, pp. 20–29, 2001.

[5] A.B. Watson and C. Ramirez, "A standard observer for spatial vision," Inv. Opht. Vis. Sci., vol. 41, no. 4, pp. S713, 2000.

[6] S. Wuerger, A. B. Watson, and A. J. Ahumada, "Toward a standard observer for spatio-chromatic detection," Proc. SPIE, vol. 4662, 2002.

[7] P. Corriveau and A. Webster, "The Video Quality Experts Group web page," http://www.vqeg.org/.

[8] A.B. Watson and J.A. Solomon, "A model of visual contrast gain control and pattern masking," J. Opt. Soc. Am. A, vol. 14, pp. 2379–2391, 1997.

[9] A.B. Watson, "Temporal sensitivity," in Handbook of Perception and Human Performance, K.R Boff et al., Eds., New York, 1986, John Wiley & Sons.