Improvement of Watson's DVQ Metric

Richard ŠIMÍK

Dept. of Radio Electronics, Czech Technical University, Technická 2, 166 27 Praha, Czech Republic

xsimik@feld.cvut.cz

Abstract. *An improvement of Watson's DVQ (Digital Video Quality) metrics is introduced. This metric was chosen for its easy implementation caused by using DCT (Discrete Cosine Transform) for video decomposition into spatial channels. The metric is upgraded by segmentation tool. This tool is used for weighting the masked differences.*

Keywords

Digital video quality metric, Watson, improvement, segmentation, DCT.

1. Introduction

A large number of digital video equipment use lossy compression of video stream. We must use it because there is an economy incentive to reduce bit rate. But lossy compression may introduce visible artifacts in video sequences and we must have an instrument for automatically evaluating their visibilities and generally the visual quality of digital video.

Recently a number of video quality metrics have been proposed, their descriptions are in [5], [6]. Possible disadvantages of these metrics are that they may have either bad model human vision or that they may require amount of memory or computation power. Watson's metric [1] used DCT transformation for video decomposition into spatial channels. It appears from DCTune metric [7] that was developed for optimization of still image compression. This metric needs lower computation requirements and it has good correspondence with subjective tests [8].

I have used segmentation since watching video you then focus only on particular areas of the scene. This focus of attention is highly scene-dependent. I propose constructing an importance map for the scene as a prediction for the focus of attention. One of the objects attracting most of attention is a human face. We will look at the human faces on the scene immediately. For segmentation we can use a robust algorithm for face detection but this step makes the model most complicated (higher computation requirements). Instead I have used a simple segmentation algorithm [3], [4] and proposed a weight coefficient for each area of the segmentation. By means of these weight coefficients I weight masked differences from the DVQ metric before their pooling.

2. Improved DVQ metric

Watson's DVQ metric [1] computes the visibility of artifacts expressed in the DCT domain. Watson considered a simple, separable model that is the product of temporal function, spatial function and orientation function:

$$T(u,v,w) = T_0\, T_w(w)\, T_f(u,v)\, T_a(u,v). \tag{1}$$

In this function T_0 is a global or minimum threshold. The remaining functions have unit peak gain – the minimum threshold is given directly by T_0.

The temporal function is the inverse of the magnitude response of a first-order discrete IIR (Infinite Impulse Response) low-pass filter (Fig. 1):

$$T_w(w) = \left| \frac{-1 + e^{\frac{1+i2\pi\tau_0 w}{\tau_0 w_s}}}{-1 + e^{\frac{1}{\tau_0 w_s}}} \right| \tag{2}$$

where w_s is sample rate in Hz and τ_0 time constant in seconds.

The spatial function (Fig. 2) is the inverse of Gaussian. The parameter f_0 corresponding to the radial frequency at which threshold is improved by a factor of e^{π}. The parameter p is the display resolution in pixels/degree and then (p/16) converts from DCT frequencies to cycles/degree (one cycle includes 8+8 pixels).

$$T_f(u,v) = e^{\pi \frac{u^2+v^2}{f_0^2}\left(\frac{p}{16}\right)^2}. \tag{3}$$

The orientation function takes into account the higher threshold for oblique frequencies and the imperfect visual summation between two component frequencies.

$$T_a(u,v) = \frac{2^{\frac{\beta-1}{\beta}}}{1 - \frac{4ru^2v^2}{\left(u^2+v^2\right)^2}} \tag{4}$$

where r and β are parameters [2].

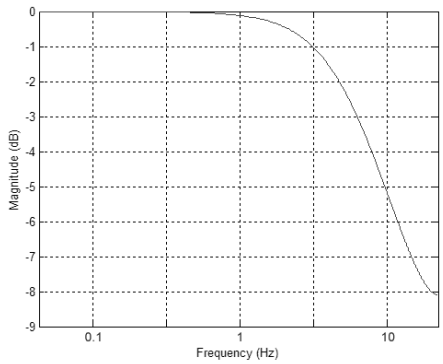

Fig. 1. The temporal function.

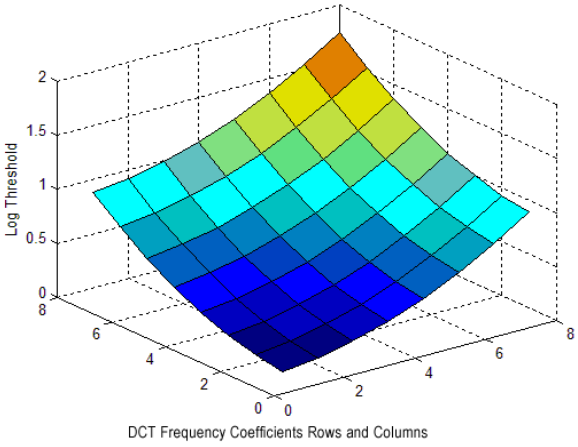

Fig. 2. The spatial function.

Fig. 3 is a block diagram of improved DVQ metric. The input of the metric is a pair of color image sequences. The first sequence is a reference and the second sequence is the test (a sequence with compression artifacts). The first step is a possible cropping to exclude regions whose quality is not of interest.

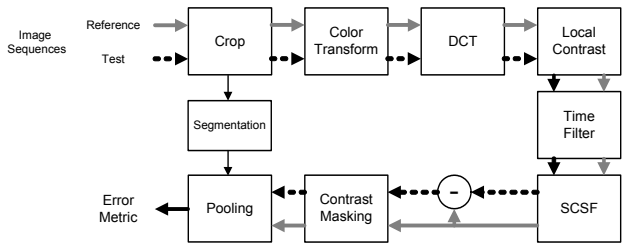

Fig. 3. The block diagram of improved DVQ metric

The next step is a transformation from the input video color format (RGB, YCbCr, …) to the color space YOZ. Watson found [1] that it could be used YCbCr color space instead of the YOZ space, too. This simplifies the color transformation for a practical use.

The next step is a video frame transformation to blocked DCT. Blocked DCT is applied to each frame in each color channel.

The next step is the transformation to Local Contrast. The Local contrast is the ratio of DCT amplitude to DC amplitude for the corresponding block.

The next step is temporal filtering which implements the temporal part of the contrast sensitivity function. The temporal filter is the second-order IIR filter, as described above. Using the IIR filter we minimize the number of video frames that must be saved in the memory. For greater simplicity we can use the first-order filter.

The DCT coefficients expressed in local contrast form are now converted to just-noticeable-differences (JNDS): We obtained respective spatial thresholds for each DCT coefficients from Equations 3 and 4. For each coefficient, we now need to determine the amount of distortion in term of JNDS units. This is done by weighting the DCT coefficients by the spatial thresholds and by computing the error at each location (the difference between the DCT coefficients in the reference and test sequence).

The DCT coefficients are then divided by their respective spatial thresholds. This implements the spatial part of the contrast sensitivity function (SCSF – Spatial Contrast Sensitivity Function). After that, the two sequences are subtracted to produce a difference sequence.

The difference sequence is then subjected to a contrast masking operation. Contrast masking is accomplished by the masking sequence that depends upon the reference sequence. The reference sequence is time-filtered by a first-order, low-pass, discrete IIR filter (with a gain of g_1 and a time constant of τ_1) after its JND conversion. These values are then raised to a power m. The values less than 1 are replaced by 1, and the result is used to divide the difference sequence. This process corresponds to the traditional contrast masking where contrasts below a threshold have no masking effect, and that above threshold the effect rises as the mth power of mask contrast in JNDS.

In another block a segmentation of the reference sequence is performed. The segmentation that is used is very simple. It segments video sequence into three types of areas: uniform areas, contours and textures [3], [4]. The input images from reference sequence are parsed pixel by pixel. The surrounding square area is considered for a given pixel, i.e. a small block is chosen, the centre of which is the considered pixel. The variance of a block is computed as well as the variance in horizontal, vertical and diagonal directions. If the variance over the block is below some predefined threshold, then the activity of the block is low and the pixel is considered to be a part of the uniform area. If there is no direction such that the ratio of variance in this direction to variance over the block is close to unity, then the pixel is considered to belong to a texture zone. If there is a direction such that this ratio is much smaller than unity, then the pixel is considered to belong to a contour whose direction is the one that yields a small variance ratio. An example of segmented image is in Fig. 4 and 5.

Fig. 4. One frame from video sequence "Claire".

Fig. 5. The segmentation of one frame from video sequence.

All points of a specific area have a weight coefficient depending on which area it falls into.

Finally the masked differences are weighted by the three coefficients from the segmentation and pooled over all dimensions {frames *f,* color channels *c,* the number of blocks in vertical and horizontal directions *by, bx,* vertical and horizontal frequencies *u, v*} to yield summary measures of visual error. This summation is done using Minkowski metric:

$$J_x = M\left(j_{f,c,by,bx,u,v}, \beta\right) = \left(\sum_x \left| j_{f,c,by,bx,u,v} \right|^{\beta}\right)^{\frac{1}{\beta}} \tag{5}$$

the exponent β has a value of 4, which is close to probability summation [4].

3. Results

Three video sequences were used for testing new metric – Claire, Carphone and Foreman. These sequences were compressed by MPEG-2 (Moving Picture Experts Group) software codec to the eight different data streams according to its bit rate. Two examples of the difference sequences for video sequence "Claire" are shown in Fig. 6 and 7.

Next the results from the metric for all test sequences with different bit rate were compared with RMSE (Root Mean Squared Error) results (Fig. 8):

Fig. 6. One frame from a difference sequence "Claire" 56kbps.

Fig. 7. One frame from a difference sequence "Claire" 200kbps.

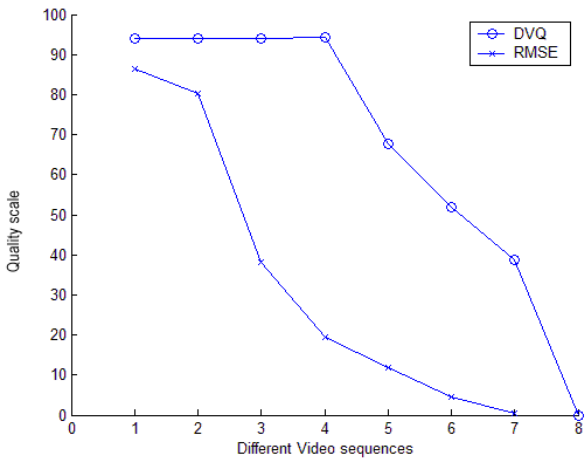

Fig. 8. A comparison between DVQ (video sequence Claire) and RMSE.

The distortion computed in Eq. (5) is a distortion measure that can be used as is. For our purpose it is expressed on the scale, defined in [10], that is modified. It is summarized in Tab. 1. The quality rating on this scale is obtained using the normalized conversion [11]:

$$Q = 100 - \frac{100}{1 + N \cdot J} \tag{6}$$

where Q is the quality rating, J is the measured distortion, and N is normalization constant (its estimation is described in [4]).

Rating	*Impairment*	*Quality*
0	Imperceptible	Excellent
25	Perceptible, not annoying	Good
50	Slightly annoying	Fair
75	Annoying	Poor
100	Very annoying	Bad

Tab. 1. Quality rating on 0 to 100 scale.

4. Conclusion

- An improvement of the digital video quality metric of Watson was proposed.
- Segmentation can improve the results from metric according to viewer recognize coding artifacts better in the uniform areas than in the textures.
- In the case of using the robust algorithm for face detection we can recognize areas those are very important for viewers and we can classify them by the relevant weight but we will consume higher computation power.

4.1 Future Tasks

- A comparison between model results and subjective tests (VQEG – Video Quality Experts Group).
- Modeling two or three temporal mechanisms. In the current model only one low-pass temporal filtering is applied but now it is believed that there is one low-pass (sustained channel) and one band-pass (transient channel) mechanism.
- Derivation of weighting coefficients for segmentation from a comparison between model results and subjective tests.

Acknowledgements

This work has been conducted at the Department of Radio electronics of the Faculty of Electrical Engineering of the Czech Technical University in Prague in the frame of the research project "Qualitative Aspects of Image Compression Methods in Multimedia Systems" and it has been supported by grant No. 102/02/0133 of the Grant Agency of the Czech Republic.

A part of this research work has been partially supported by the research program No. MSM 212300014 "Research in the Area of Information Technologies and Communications" of the Czech Technical University in Prague.

References

[1] WATSON, A. B. Toward a perceptual video quality metric. In *Proc. SPIE.* San Jose, CA, 1998, vol. 3299, p. 139–147.

[2] PETERSON, H., AHUMADA, A. J., WATSON, A. An Improved Detection Model for DCT Coefficient Quantization. In *Proc. SPIE.* 1993, vol. 1913, p. 191-201.

[3] EGGER, O., LI, W., KUNT, M. High Compression Image Coding Using an Adaptive Morphological Subband Decomposition. In *Proceedings of the IEEE, Special Issue on Advances in Image and Video Compression*, 1995, vol. 83, no. 2, p. 272-287.

[4] van den BRANDEN LAMBRECHT, Ch. J. Perceptual Models and Architectures for Video Coding Applications. *Ph.D. Thesis*, Ecole Polytechnique Federale de Lousanne, Lausanne, EPFL, 1996.

[5] NADENAU, M. J., WINKLER, S. et al. Human Vision Models for Perceptually Optimized Image Processing – A Review. In *Proc. of the IEEE.* 2000.

[6] WINKLER, S. Vision Models and Quality Metrics for Image Processing Applications. *Ph.D. Thesis*, Ecole Polytechnique Federale de Lousanne, Lausanne, EPFL, 2000.

[7] WATSON, A. B. DCT quantization matrices visually optimized for individual images. In *Proc. SPIE.* 1993, vol. 1913, p. 202–216.

[8] WATSON, A. B. et al. Design and performance of a digital video quality metric. In *Proc. SPIE.* San Jose, CA, 1999, vol. 3644, p. 168–174.

[9] FREDERICKSEN, R. E., HESS, R. F. Estimating multiple temporal mechanisms in human vision. In *Vision Research.* 1998, vol. 7, no. 38, p. 1023–1040.

[10] CCIR. Method for the Subjective Assessment of the Quality of Television Pictures. In *13th Plenary Assembly, Recommendation 500.* 1974, vol. 11, p. 65-68.

[11] COMES, S., MACQ, B. Human Vision Quality Criterion. In *SPIE Visual Communications and Image Processing.* 1990, vol .1360, p. 2-7.

About Author...

Richard ŠIMÍK was born in 1976 in Sušice (Czech Republic). He received the Ing. (M.Sc.) degree in electrical engineering from the Faculty of Electrical Engineering, Czech Technical University in Prague in 2000. At present he is a Ph.D. student at the Department of Radio electronics, Faculty of Electrical Engineering, Czech Technical University in Prague.