FUNQUE: FUSION OF UNIFIED QUALITY EVALUATORS

Abhinau K. Venkataramanan[⋆] *Cosmin Stejerean*[†] *Alan C. Bovik*[⋆]

[⋆] The University of Texas at Austin, Austin, TX 78705 USA
[†] Meta Platforms, Inc., Menlo Park, CA 94025 USA

ABSTRACT

Fusion-based quality assessment has emerged as a powerful method for developing high-performance quality models from quality models that individually achieve lower performances. A prominent example of such an algorithm is VMAF, which has been widely adopted as an industry standard for video quality prediction along with SSIM. In addition to advancing the state-of-the-art, it is imperative to alleviate the computational burden presented by the use of a heterogeneous set of quality models. In this paper, we unify "atom" quality models by computing them on a common transform domain that accounts for the Human Visual System, and we propose FUNQUE, a quality model that fuses unified quality evaluators. We demonstrate that in comparison to the state-of-the-art, FUNQUE offers significant improvements in both correlation against subjective scores and efficiency, due to computation sharing.

Index Terms— Video Quality Assessment, Human Visual System, Visual Multimethod Assessment Fusion

1. INTRODUCTION

The share of video in mobile traffic is expected to reach 82% by the year 2022 [1]. Owing to this explosion of videos online, Video Quality Assessment (VQA) has emerged as a key area of research. While the most reliable form of VQA is subjective VQA, where videos are rated by human subjects, practical VQA relies on objective VQA models. In the case of streaming and Video On Demand (VOD) applications, the pristine video is available for use as a reference against which distorted videos may be evaluated. Therefore, Full-Reference (FR) VQA algorithms are of special interest.

The Video Multimethod Assessment Fusion (VMAF) [2] quality model has been widely adopted as an industry standard for quality assessment of compressed videos. In particular, VMAF has found use in the perceptual optimization of encoding recipes, comparison of codecs [3, 4], and to evaluate enhancement/precoding methods [5]. The current widely adopted version of the VMAF quality model is VMAF v0.6.1, which uses Spatial Visual Information Fidelity (VIF) [6] at 4 scales, the Detail Loss Metric (DLM, called ADM in VMAF) [7], and Temporal Difference (TD, called Motion in VMAF), as its "atom features," which are fused using a support vector regressor (SVR).

Notably, the VIF model [6] was originally defined in the steerable-pyramid wavelet domain [8], and has been reformulated using dyadic wavelets [9]. Thus, the wavelet domain is the "natural domain" for VIF. On the other hand, the spatial VIF model is a version of VIF that has been optimized for speed by omitting the wavelet decomposition, at the cost of a lower correlation against subjective scores. However, since the computation of DLM requires an expensive 4-level wavelet transform, sharing the wavelet decomposition will allow the computation of VIF in its natural domain, while improving computational efficiency. This notion of using a common decomposition to "unify" the atom features is a foundational idea of this work.

2. RELATED WORK

Since VMAF is applied on the luma channel and the only temporal feature is the temporal difference of the reference video, attempts at improving VMAF have typically focused on the inclusion of color and temporal features. ColorVMAF [10] introduces color information by computing features on the chroma channels, while the use of spatio-temporal features has been explored by Ensemble VMAF [11]. The robustness of VMAF to video enhancement has been improved with the development of Anti-Hacking VMAF [12].

The Enhanced VMAF model (EVMAF) [13] is the most

Table 1. Design choices considered for FUNQUE

Design Choice	Options
Wavelet	Haar / Db2
Wavelet Levels	1 - 4
CSF	Frequency filter / Spatial filter / Li SW / Watson SW
CSF Sharing	Yes / No
SAST	Yes / No

This research was sponsored by a grant from Facebook Video Infrastructure, and by grant number 2019844 for the National Science Foundation AI Institute for Foundations of Machine Learning (IFML).

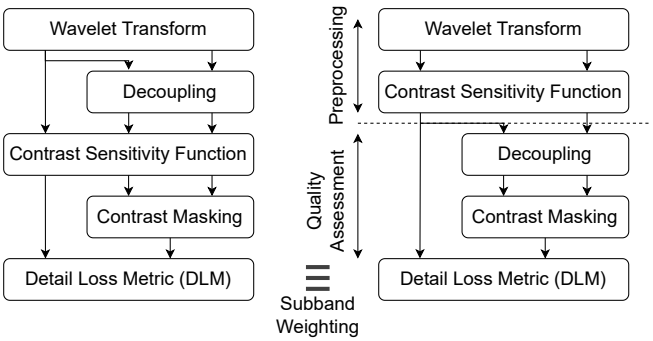

Fig. 1. An equivalent reformulation of DLM

recent attempt at improving the VMAF model. Similar to Ensemble VMAF, EVMAF combines two models, one trained on a private database, and one on a public database. However, EVMAF incorporates motion using a dynamic texture feature, and expands the feature set significantly, using greedy feature selection to obtain the final models. In addition to VMAF v0.6.1, the EVMAF model is used as a baseline against which the performance of FUNQUE is compared.

3. ALGORITHM

The foundation of the FUNQUE framework is the use of a unified transform that is shared by all "atom" quality models. This unified transform consists of a wavelet transform and HVS processing using a model of the contrast sensitivity function (CSF) and is described in detail in Section 3.1. The following four "categories" of atom features have been considered. The feature selection method used to select the final feature set has been described in Section 4.

1. DLM
2. Structural Similarity (SSIM) [14] or Enhanced SSIM (ESSIM) [15] computed in the wavelet domain (WD-SSIM and WD-ESSIM), as described in Section 3.2.
3. VIF using vector and scalar Gaussian Scale Mixture models [6], VIF-Edge and VIF-Approx features [9], and VIF-Scale features, which consist of applying scalar VIF on low-pass subbands at each wavelet level, similar to VMAF's VIF features.
4. Motion, analogous to VMAF's TD feature, computed as the Mean Absolute Difference between low-pass subbands of successive frames of the reference video.

3.1. Unified Transform

The DLM algorithm is illustrated in Figure 1. The CSF is applied in DLM by multiplying each subband by a CSF value. We will refer to any such mechanism that assigns weights to each subband as a "subband-weighting" (SW) mechanism, and we will refer to the method used in [7] as "Li subband weighting" (Li SW). On the other hand, the ADM feature used in VMAF is a version of DLM that uses a different set of weights, obtained from [21]. We will refer to this method as "Watson subband weighting" (Watson SW).

A closer look at the Decoupling step of DLM reveals that the order of Decoupling and any SW CSF may be reversed. This allows the reinterpretation of DLM's workflow as quality assessment performed on an HVS-aware transform. This raises the possibility of sharing the HVS-aware transform among all the atom features. Indeed, our experiments revealed that CSF sharing leads to a significant boost in performance. Furthermore, since SW is a "coarse" CSF method, due to the use of uniform weights within subbands, finer means of applying the CSF may be considered.

The frequency domain is a natural choice for applying the CSF since it is a function of spatial frequency. Consequently, the following frequency-domain model of the CSF was used [22], which was also used to derive Li's subband weights.

$$\text{CSF}(f) = (0.31 + 0.69f)e^{-0.29f}, \tag{1}$$

where f has the units cycles/degree. The CSF is applied independently on horizontal and vertical frequencies.

An equivalent "continuous-angle" filter in the spatial domain may be obtained using the Inverse Fourier Transform. For practical use, the continuous filter is sampled and truncated to obtain a 21-tap filter that is applied separably in 2D to perform CSF filtering in the spatial domain. The effectiveness of spatial CSF filtering is demonstrated in Section 5.1.

Table 2. Databases used for model selection

Database	Size	Codec(s)
CC-HDDO [3]	90	HM, AV1
BVI-HD [16]	192	HM
CC-HD [4]	108	HM, AV1, VTM
IVP [17]	100	Dirac, JM, MPEG-2
MCL-V [18]	96	x264
Netflix Public [2]	70	x264
SHVC [19]	64	HM
VQEGHD3 [20]	72	MPEG-2, JM

Table 3. The proposed FUNQUE model

Atom Features	Wavelet (Levels)	CSF	SAST
WD-ESSIM + VIF-Scales 1 & 2 + DLM + Motion	Haar (1)	Spatial	Yes

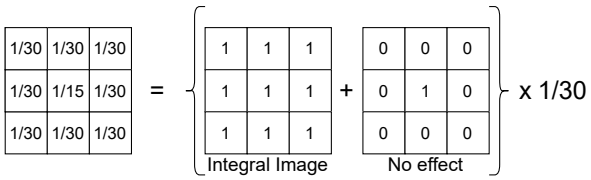

Fig. 2. Decomposing DLM's contrast masking kernel

3.2. Wavelet-Domain Structural Similarity

The ESSIM model [15] improves upon SSIM by using small, strided rectangular windows to compute local quality scores, Coefficient of Variation (CoV) pooling for spatial aggregation, and the Self-Adaptive Scale Transform (SAST) to rescale frames before quality assessment. However, since SAST must be applied commonly to all atom features, its use is investigated as a global design choice in Section 4. So, in this paper, we consider CoV-pooling to be the defining factor of ESSIM, and we consider SSIM to be any mean-pooled version of the algorithm.

In this section, we describe a method for computing SSIM and ESSIM directly from Haar wavelet subband coefficients, which allows us to use the unified transform described in Section 3.1. This is achieved by leveraging the orthonormality of the Haar bases, and the structure of the Haar transform.

Using these properties, local statistics within disjoint blocks of size $2^L \times 2^L$ may be obtained directly from the wavelet coefficients. For simplicity, consider a pair of images x, y of size $M \times N$ such that both M and N are divisible by 2^L. Now consider their L-level wavelet decompositions. Let $H_{x,k}, V_{x,k}, D_{x,k}$, and $H_{y,k}, V_{y,k}, D_{y,k}$ denote the horizontal, vertical, and diagonal subbands at level k of their wavelet decompositions respectively, and $A_{x,L}$, $A_{y,L}$ be the respective residual low pass (approximation) subbands. Then, local means, variances, and covariances may be obtained as

$$\mu_L(i,j) = 2^{-L} A_L(i,j), \tag{2}$$

$$\sigma_L^2(i,j) = 2^{-2L} \sum_{k=1}^{L} \sum_{P_{ij}^k} \sum_{\{H,V,D\}} C_k^2(m,n), \tag{3}$$

$$\sigma_{xy,L}(i,j) = 2^{-2L} \sum_{k=1}^{L} \sum_{P_{ij}^k} \sum_{\{H,V,D\}} C_{x,k}(m,n) C_{y,k}(m,n), \tag{4}$$

where $P_{ij}^k = \{(m,n) \mid i2^{L-k} \leq m < (i+1)2^{L-k}, j2^{L-k} \leq n < (j+1)2^{L-k}\}$, and $C \in \{H,V,D\}$ denotes a subband. These local statistics may be used to compute both WD-SSIM and WD-ESSIM, as in [15].

3.3. Integral Image-based Optimization

In addition to the use of a unified transform and computing SSIM directly from wavelet coefficients, integral images [23] have been used to effectively minimize the number of convolution operations. Specifically, we have optimized the computation of local statistics for VIF using integral images, instead of convolution, as in [15]. Furthermore, the non-separable 3×3 kernel used in DLM has been decomposed as shown in Figure 2. Local sums are then computed using integral images, and the Kronecker delta leaves the image unchanged. Hence, the only convolution in DLM is eliminated.

4. EXPERIMENTS

The framework presented in Section 3 offers a few free design choices. Specifically, let a "configuration" correspond to a choice of wavelet, number of wavelet levels, CSF method, whether to share the CSF, and whether to apply SAST. Note that spatial and frequency-domain CSFs must be shared since they are applied before the wavelet decomposition. Table 1 lists all the design choices considered in our experiments, which led to a total of 96 configurations.

To identify the best configuration, we conducted experiments using the set of 8 databases listed in Table 2. These databases were chosen since they represent the popular use-case of video compression, and they were used to develop EVMAF. In order to avoid large models that use several features of the same type, we performed feature selection under the constraint that at most one feature is selected from each category described in Section 3. An exhaustive search was performed to optimize the Spearman Rank Order Correlation Coefficient (SROCC) during cross-validation over 5000 80-20 splits of the CC-HDDO database. The best model so obtained was tested on the other 7 databases, and the average test SROCC was obtained using Fisher averaging [13].

5. RESULTS

The best FUNQUE model has been described in Table 3, and Table 4 summarizes the correlations against subjective scores achieved by VMAF, EVMAF, and FUNQUE. Since FUNQUE was trained only on the CC-HDDO database, we retrained VMAF v0.6.1 for a fair comparison. Indeed, the retrained model achieved a significantly lower performance, demonstrating the effectiveness of the private database in training. Similarly, we consider EVMAF M2 to be the primary baseline since EVMAF M1, and therefore, the combined EVMAF model, was trained on a private database. The best model on each database among FUNQUE and the "fair" baselines has been highlighted in bold. In addition, any better performing "unfair" baseline has also been highlighted.

From Table 4, it may be observed that FUNQUE significantly outperforms retrained VMAF v0.6.1, and also outperforms the high-complexity EVMAF M2 model. In addition, despite training only on public data, FUNQUE outperforms VMAF v0.6.1 off-the-shelf and rivals the performance of EVMAF M1, both of which were trained on private Netflix data.

Table 4. Comparison of FUNQUE's performance with baseline fusion models

Model	BVI-HD	CC-HD	IVP	MCL-V	NFLX-P	SHVC	VQEGHD3	Average
VMAF v0.6.1	**0.7962**	0.8723	0.8786	0.7766	0.9104	0.8442	**0.9114**	0.8631
Retrained VMAF v0.6.1	0.7516	**0.8920**	0.7156	0.8133	0.8756	0.7205	0.7692	0.8019
Enhanced VMAF - M1	0.8067	0.8595	0.9060	0.8044	0.9168	0.8652	0.9221	0.8761
Enhanced VMAF - M2	0.7920	0.8376	0.8810	**0.8327**	0.9141	0.8591	0.8729	0.8600
Enhanced VMAF	0.8057	0.8783	0.9022	0.8282	0.9253	**0.8796**	0.9241	**0.8842**
FUNQUE	**0.7959**	0.8315	**0.9186**	0.7302	**0.9358**	**0.8769**	**0.9088**	**0.8715**

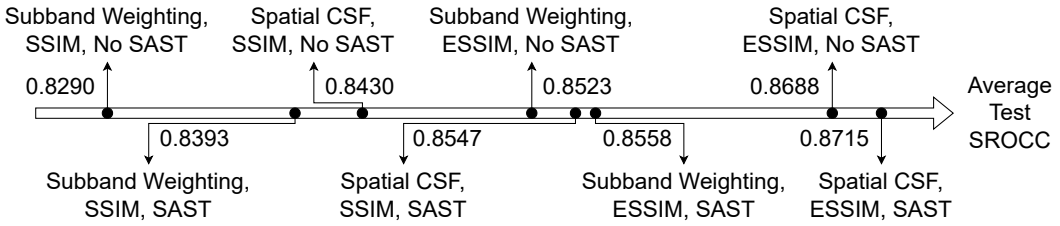

Fig. 3. Visualizing the effect of spatial CSF filtering, Enhanced SSIM, and SAST on performance.

5.1. Ablation Study

In order to understand FUNQUE's superior performance, we would like to investigate three key design choices - the use of ESSIM vs. SSIM, the use of SAST, and the use of the spatial CSF vs. SW CSFs. Since our experiments revealed that not sharing CSF decreased performance significantly, this choice has been omitted from the ablation study.

The performances of the eight models so obtained have been illustrated in Figure 3. From the figure, it may be observed that all eight models outperform retrained VMAF v0.6.1, which demonstrates the impact of CSF sharing. Secondly, both ESSIM and Spatial CSF contribute roughly equally to FUNQUE's performance. In other words, despite omitting the expensive 21-tap CSF filter, about 50% of the reported improvement may be achieved. Finally, we observe that using SAST always improves model performance. Since SAST effectively scales images to half the original resolution, it also improves efficiency.

5.2. Timing Analysis

In order to highlight the computational benefits of FUNQUE, estimates of the number of operations per pixel (OPP) required to compute VMAF and FUNQUE were obtained. The ratio of estimated OPPs is reported as the expected speedup. Furthermore, a practical estimate of the speedup was obtained by measuring the ratio of the average running time of the two models on ten videos from the Netflix-Public database.

Table 5. Timing analysis of FUNQUE models

Model	Runtime	Ops Per Pixel	Observed Speedup	Expected Speedup
PyVMAF	105.23 (s)	219.61	1	1
FUNQUE	**12.73 (s)**	**39.30**	**8.265**	**5.588**

Since FUNQUE was implemented in Python, the VMAF model was reimplemented in Python for a fair comparison. We refer to this model as PyVMAF. From Table 5, it may be observed that FUNQUE reports a significant speedup of over 8$\times$ as compared to PyVMAF! Since EVMAF uses more features and requires optical flow estimation, it would already be much slower than VMAF v0.6.1. Therefore, it has not been included in this timing analysis.

6. CONCLUSION

In summary, we have proposed a framework to unify quality evaluators by computing them from a common HVS-sensitive transform and fusing them using an SVR. FUNQUE significantly outperforms both the baseline models, i.e., VMAF v0.6.1 and Enhanced VMAF, at less than 1/8th of the computational cost. An open-source implementation of FUNQUE is available at `https://github.com/utlive/funque`.

In the future, we see merit in including more sophisticated color and motion-sensitive features, as in ColorVMAF [10] and EVMAF. A more extensive feature set may be considered too, with a special focus on wavelet-domain features. Finally, more sophisticated models of the CSF, and even the HVS, may be considered.

7. REFERENCES

[1] T. Barnett, S. Jain, U. Andra, and T. Khurana, "Cisco visual networking index (VNI) complete forecast update, 2017-2022," *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*, 2018.

[2] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, p. 2, 2016.

[3] A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull, "A subjective comparison of AV1 and HEVC for adaptive video streaming," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4145–4149.

[4] F. Zhang, A. V. Katsenou, M. Afonso, G. Dimitrov, and D. R. Bull, "Comparing VVC, HEVC and AV1 using objective and subjective assessments," *ArXiv*, vol. abs/2003.10282, 2020.

[5] E. Bourtsoulatze, A. Chadha, I. Fadeev, V. Giotsas, and Y. Andreopoulos, "Deep video precoding," *ArXiv*, vol. abs/1908.00812, 2019.

[6] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[7] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[8] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings., International Conference on Image Processing*, vol. 3. IEEE, 1995, pp. 444–447.

[9] S. Rezazadeh and S. Coulombe, "A novel discrete wavelet transform framework for full reference image quality assessment," *Signal, Image and Video Processing*, vol. 7, no. 3, pp. 559–573, 2013.

[10] L.-H. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik, "Perceptual video quality prediction emphasizing chroma distortions," *IEEE Transactions on Image Processing*, vol. 30, pp. 1408–1422, 2021.

[11] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2019.

[12] Z. Li, K. Swanson, C. Bampis, L. Krasula, and A. Aaron, "Toward a better quality metric for the video community," *The Netflix Tech Blog*, p. 2, 2020.

[13] F. Zhang, A. Katsenou, C. Bampis, L. Krasula, Z. Li, and D. Bull, "Enhancing VMAF through new feature integration and model combination," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[14] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[15] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid, "A hitchhiker's guide to structural similarity," *IEEE Access*, vol. 9, pp. 28 872–28 896, 2021.

[16] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[17] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. (2009) IVP subjective quality video database. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/

[18] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 1–9, 2015.

[19] Y. He, Y. Ye, F. Hendry, Y. K. Wang, and V. Baroncini, "SHVC verification test results," *JCT-VC Meeting*, no. JCTVC-W0095, 2016.

[20] (2010) Report on the validation of video quality models for high definition video content. Video Quality Experts Group. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv.aspx

[21] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1164–1175, 1997.

[22] K. N. Ngan, K. S. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1743–1750, 1989.

[23] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.