Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

Article

Hierarchical Gradient Similarity Based Video Quality Assessment Metric

Jie Yang [1], Jian Xiong [1]*, Guan Gui [1], Rongfang Song [1], Wang Luo [2], and Xianzhong Long [3]

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, 210003; jyang@njupt.edu.cn (J.Y.), jxiong@njupt.edu.cn (J.X.), guiguan@njupt.edu.cn (G.G.), songrf@njupt.edu.cn (R.-F.S.)
[2] Nari Group Corporation (State Grid Electric Power Research Institute), Nanjing, China; luowang@sgepri.sgcc.com.cn (W.L.)
[3] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China, 210003; lxz@njupt.edu.cn (X.-Z.L.)

* Correspondence: jxiong@njupt.edu.cn

Abstract: Video quality assessment (VQA) plays an important role in video applications for quality evaluation and resource allocation. It aims to evaluate the video quality consistent with the human perception. In this letter, a hierarchical gradient similarity based VQA metric is proposed inspired by the structure of the primate visual cortex, in which visual information is processed through sequential visual areas. These areas are modeled with the corresponding measures to evaluate the overall perceptual quality. Experimental results on the LIVE database show that the proposed VQA metric significantly outperforms the state-of-the-art VQA metrics.

Keywords: hierarchical video quality assessment; human visual systems; primate visual cortex; full reference

0. Introduction

Recently, video quality assessment (VQA) metrics which can evaluate the video quality consistent with the human perception have received increased attention. VQA metrics are generally classified into three categories, full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics. A full-reference (FR) metric aims to evaluate the qualities of distorted videos with the full available reference videos. Peak signal-to-noise ratio (PSNR) and Mean square error (MSE) [1,2] are the most widely used FR metrics. These indices are simple to be calculated and convenient to be adopted. But they show poor consistency with the subjective evaluations [3].

Many efforts have been made to investigate the FR VQA algorithms. Structural similarity index (SSIM) [4] is the most popular metric. The comparison functions of luminance, contrast and structure are designed and combined to obtain the overall quality. SSIM based VQA metrics have been proposed by introducing motion information and temporal weighting schemes [5,6]. These metrics are developed based on the assumption that the degradation of perceptual qualities is highly related to the change of the structural information. Moreover, gradient based metrics have been proposed to describe the loss of the structural information [7,8]. In [8], edge-strength similarities were calculated for all pixels to acquire the overall quality score for each frame. In [9], the gradient based 3-D structure tensors were decomposed to evaluate the video perceptual quality. Spatio-temporal gradient features were extracted to derive the 3-D structure tensor, and the corresponding eigenvalues and eigenvectors were used to evaluate the video perceptual quality.

© 2017 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

In this letter, a VQA metric is designed based on a hierarchical gradient similarity model. This model is inspired by functional principles of the processing hierarchies in the primate visual system [10], which is characterized by a sequence of visual areas. These areas are modeled by hierarchical gradient measures to evaluate the score of each frame. The visual attention similarity evaluated by an efficient measure is also involved in the proposed metric. Then, an averaging pooling is performed to obtain the final score of the video sequence. Experimental results show that the proposed VQA metric outperforms the state-of-the-art VQA metrics.

1. Hierarchical Video Quality Assessment

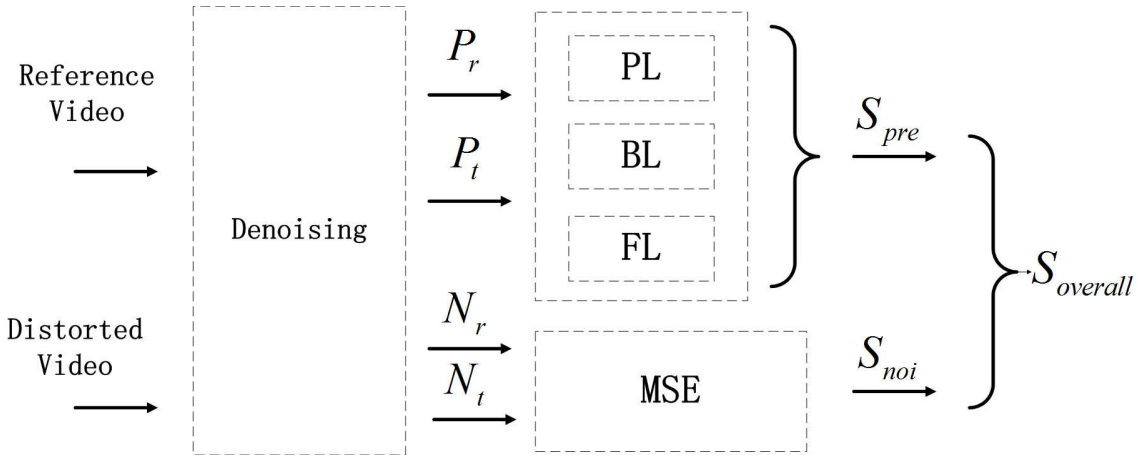

Figure 1. Flowchart of the proposed model.

As shown in Fig. 1, the neuronal processing of visual information starts from the retina. Before the visual information reaches the visual cortex, it projects to a visual area named lateral geniculate nucleus (LGN). This stage is called as *precortical processing* [10]. The occipital part of the primary visual cortex covers area V1-V4 and MT [11]. In the early visual areas, simple image features are extracted over small local regions. Then this information is transmitted to the higher visual areas, in which more complex features are extracted covering larger and larger regions. The occipital part gives input to the *ventral pathway* (VP) and *dorsal pathway* (DP) [11]. The ventral pathways (V1→ V2→ V4) is critical for objection discrimination and the dorsal pathway (V1→ V2→ MT) is functionally related to visual motion [10]. In this study, in order to assess the video quality consistent with the human perception, the assessment of visual information is inspired by the processing hierarchies in the primate visual system. LGN is modeled as band-pass filtering, and the two pathways are modeled with hierarchical gradient measures. Furthermore, visual attention similarity evaluated by an new efficient measure is involved successively.

1.1. Modeling of the Precortical Processing

The precortical processing stage in the area LGN has a band-pass filtering characteristic for luminance stimuli [10]. This can be modeled by a denoising operation, such as VBM3D [12]. Recent study [7] also showed that the perceptual distortions can be classified into content-dependent distortions and content-independent distortions. The content-independent distortions are mainly related to the additive noise. Thus, using the denoising operation, both of the reference and distorted frames are decoupled into two parts, the *prediction part* and the *noise part*. Since MSE presents a good match with the additive noise [13], the MSE is adopted to evaluate the degradation of the noise part:

$$S_{noi}(\mathbf{N}_r, \mathbf{N}_t) = 1 - \frac{\log_{10}(1 + \mathrm{MSE}(\mathbf{N}_r, \mathbf{N}_t))}{\log_{10}(255^2)}, \tag{1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

where $\mathbf{N}_r$ and $\mathbf{N}_t$ are the noise part of the reference and test videos, respectively; $\mathrm{MSE}(\mathbf{N}_r, \mathbf{N}_t)$ denotes the MSE between $\mathbf{N}_r$ and $\mathbf{N}_t$. The denominator $\log_{10}(255^2)$ is used to normalize the metric into the range [0,1]. In the numerator, adding 1 is to avoid to be smaller than 0. A S_{noi} approximates to 1more means the distortion is more weak.

1.2. Modeling the Dorsal Pathway

The areas V1 and V2 in the dorsal pathway contain cells that respond preferentially to linear oriented patterns, such as the edges, bars, and gratings [10]. Edge detection such as Sobel filter can be used to model the processing of these areas. Furthermore, the area MT is dedicated to visual motion such as motion gradients, motion-defined edges, locally opposite motions. It can be modeled as the spatio-temporal gradient.

Since the spatio-temporal gradient vector also contains the components of the spatial version, the similarity of the spatio-temporal gradient vector is used to model the dorsal pathway. To balance the effect of the temporal and spatial gradients, each component is divided by the sum of positive filter coefficients, respectively. The similarity in the dorsal pathway is

$$S_{dp}(x_r, x_t) = \frac{2\|\mathbf{g}^r\|_2\|\mathbf{g}^t\|_2 + C_1}{(\mathbf{g}^r)^2 + (\mathbf{g}^t)^2 + C_1} \cdot \frac{\mathbf{g}^r \cdot \mathbf{g}^t + C_1}{\|\mathbf{g}^r\|_2\|\mathbf{g}^t\|_2 + C_1}, \tag{2}$$

where $S_{dp}(x_r, x_t)$ denotes the gradient similarity between x_r and x_t, which are the pixels in prediction parts of the reference frame and the distorted frame, respectively. The vectors $\mathbf{g}^r$ and $\mathbf{g}^t$ denote the corresponding spatio-temporal gradient vectors, which are calculated by the Sobel filter along x, y and t directions, respectively, i.e., $\mathbf{g} = (g_x, g_y, g_t)$. The Sobel kernel for the t direction is a $3 \times 3 \times 3$ matrix [9]. The parameter C_1 is a small constant to avoid the denominator being zero, and is set as $C_1 = 0.03 \times 255^2$. The first term represents the similarity of the strengthes between $\mathbf{g}^r$ and $\mathbf{g}^t$. The second term represents the similarity of the directions between the two gradient vectors. Eqn (2) can be further simplified to:

$$S_{dp}(x_r, x_t) = \frac{2\mathbf{g}^r \cdot \mathbf{g}^t + C_1}{(\mathbf{g}^r)^2 + (\mathbf{g}^t)^2 + C_1}. \tag{3}$$

Using (3), each pixel will get the DP similarity.

1.3. Modeling the Ventral Pathway

Both of the ventral and dorsal pathways contain the areas V1 and V2. Therefore, in order to reduce the repetitive computation, only the area V4 should be modeling in the ventral pathway. The area V4 is important for the perception of shape/curvature discrimination. The features are extracted over larger regions instead of the local regions. In this study, it is modeled by the block-level gradient vectors similarities. The reference and distorted video frames are split into 8x8 non-overlapped blocks. The mean values of the blocks construct a down-sampled versions of the images. The spatial gradient of the down-sampled images are used to evaluate the similarity in the ventral pathway:

$$S_{vp}(b_r, b_t) = \frac{2\mathbf{g}_b^r \cdot \mathbf{g}_b^t + C_1}{(\mathbf{g}_b^r)^2 + (\mathbf{g}_b^t)^2 + C_1}, \tag{4}$$

where the $S_b(b_r, b_t)$ denotes the block-level gradient similarity between the blocks b_r and b_t, which are the blocks in the prediction parts of the reference frame and the distorted frame, respectively. The formula is similar to Eqn (3), whereas the different is the vectors $\mathbf{g}_b^r$ and $\mathbf{g}_b^t$ are the 2-D spatial gradient vectors of the down-sampled images. Using (4), each block will get the VP similarity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

1.4. Visual Attention Similarity

Representations in the visual cortex are known to be overcomplete. Visual attention models [14,15] show that the human visual system is more sensitive to the salient regions. The similarities of only the salient pixels are selected to evaluate the perceptual quality. Similar to [9], pixels are determined to be the salient pixels if their spatio-temporal gradient magnitudes are above a threshold in either the reference video or the distorted video. The threshold is defined to be the average of the k^{th} largest gradient magnitudes in the prediction parts of the reference frame and the distorted frame, respectively. We denote the set of the salient pixels in P_r and P_t as $\mathbf{C}_r$ and $\mathbf{C}_t$, respectively. The union of $\mathbf{C}_r$ and $\mathbf{C}_t$, denoted as $\mathbf{C}_r \bigcup \mathbf{C}_t$, is the set of the salient pixels selected to be processed.

Furthermore, the averaging pooling on the similarities of the salient pixels can be used to evaluate overall similarity. However, it may lose the changes of the visual attention, and cannot represent the degradation of the whole frame efficiently. Therefore, the similarity of the visual attention is introduced as

$$S_{va}(\mathbf{P}_r, \mathbf{P}_t) = \frac{|\mathbf{C}_r|}{|\mathbf{C}_r \bigcup \mathbf{C}_t|}, \tag{5}$$

where $S_{va}(\mathbf{P}_r, \mathbf{P}_t)$ denotes the attention similarity between the prediction parts $\mathbf{P}_r$ and $\mathbf{P}_t$ which are decoupled from the reference frame and the distorted frame, respectively. The numerator $|\mathbf{C}_r|$ denotes the number of the salient pixels in the reference frame. The denominator $|\mathbf{C}_r \bigcup \mathbf{C}_t|$ denotes the number of the salient pixels in the union set. The difference between the denominator and the numerator represents the newly increased salient pixels. Thus, the ratio between the denominator and the numerator represents the VA similarities.

1.5. Overall Score

In the above subsections, degradations in different visual areas are modeled with the corresponding similarities. The final quality index of each frame can be calculated by combining these similarities, as

$$S_{pre}(\mathbf{P}_r, \mathbf{P}_t) = S_{va} \cdot \underset{x \in \{\mathbf{C}_r \bigcup \mathbf{C}_t\}}{\text{Avg}} S_{dp}(x) \cdot S_{vp}(x), \tag{6}$$

where $S_{pre}(\mathbf{P}_r, \mathbf{P}_t)$ denotes the quality score of the prediction part which is decoupled from the distorted frame $\mathbf{F}_t$, and the parameter x denotes the salient pixel. That is, the terms $S_{dp}(x)$ and $S_{vp}(x)$ denote the similarity of the pixel x in the dorsal and ventral pathway, respectively. The term S_{va} denotes the visual attention similarity. As in [7], the overall quality score of the frame is calculated as,

$$S_{overall}(\mathbf{F}_r, \mathbf{F}_t) = (S_{pre}(\mathbf{P}_r, \mathbf{P}_t))^{S_{noi}(\mathbf{N}_r, \mathbf{N}_t)}, \tag{7}$$

where S_{pre} and S_{noi} denote the quality score of the prediction part and the noise part, respectively. Finally, all of the frame scores are averaged to give the final video quality index.

2. Experimental Results

The effectiveness of the proposed VQA metrics is evaluated by the consistency between the objective scores and the subjective scores (DMOS). The consistency is measured by the Pearson correlation coefficient (PCC) and the Spearman rank order correlation coefficient (SROCC). The LIVE subjective quality video database [18] is used to evaluate the performance of the proposed VQA metric. The parameter k is set to $0.35 \times W \times H$ through exhaustive experiments, where W and H are the width and the height of the video sequence. Results with the state-of-the-art VQA metrics, including PSNR, SW-SSIM [5], MC-SSIM [6], MOVIE [17], STSI [9], VQM [16] are compared. The results of PSNR and Picture Quality Analyzer are quoted from [9].

Table 1 shows the PCC and SROCC of each metric on the LIVE database. It is observed that the proposed hierarchical VQA (HVQA) metric significantly outperforms all of the other metrics according to both the two indicators. The gradient similarity based VQA metrics (HVQA and STSI)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

Table 1. PERFORMANCE COMPARISON ON THE LIVE DATABASE

Methods	Pearson CC	Spearman CC
VQM [16]	0.702	0.723
MOVIE [17]	0.786	0.810
STSI [9]	0.779	0.778
SW-SSIM [5]	0.585	0.596
MC-SSIM [6]	0.679	0.698
PSNR [9]	0.368	0.404
PQR (by PQA500) [9]	0.695	0.712
DMOS (by PQA500) [9]	0.695	0.711
Proposed (DP)	0.775	0.769
Proposed (VP)	0.736	0.740
Proposed (VA)	0.759	0.761
Proposed (DP&VP)	0.810	0.807
Proposed (VP&VA)	0.804	0.811
Proposed (DP&VA)	0.817	0.816
Proposed (HVQA)	**0.832**	**0.833**

Table 2. PCC SCORES OF VQA METRICS ON EACH KIND OF DISTORTION IN LIVE DATABASE

Methods	Wireless	IP	H.264	MPEG2
PSNR [9]	0.4675	0.4108	0.4385	0.3856
VQM [16]	0.7325	0.6480	0.6459	**0.7860**
STSI [9]	0.7544	0.8072	0.8298	0.6624
SW-SSIM [5]	0.5867	0.5587	0.7206	0.6270
PQR (PQA500) [9]	0.6464	0.7300	0.7455	0.6456
DMOS (PQA500) [9]	0.6426	0.7295	0.7427	0.6445
Proposed (HVQA)	**0.8109**	**0.8264**	**0.8445**	0.7654

perform better than SSIM based metrics such as SW-SSIM, and MC-SSIM. It indicates that the change of the edge gradient is highly related to the degradation of the perceptual visual quality. This is reasonable for the areas V1 and V2 are sensitive to the edge patterns. However, the proposed HVQA metric performs significantly better than STSI. The reason is that degradations over the large regions, such as the packet-loss on the flat regions, cannot be represented efficiently by only the pixel-level gradient similarities. The similarities of the area V4 and visual attention can improve the efficiency of the metrics. Furthermore, the proposed metric is compared with the optical flow based VQA metric, MOVIE. HVQA significantly outperforms the MOVIE index (SROCC increment: 0. 046), which performs the best in all of the comparison metrics.

To evaluate the effectiveness of the gradient similarities at each visual area, different combinations are reported in Table 1. The combinations can be classified into 3 categories. The category I methods use only one of the 3 measures, including DP, VP, VA. The category II methods use two of the 3 measures, including DP&VP, VP&VA, DP&VA. The category III method uses all of the 3 measures, i.e., HVQA. It is observed that the average SROCC and PCC of the category I methods are 0.757. Thus, the similarities of single measure are efficient to represent the degradation of perceptual visual quality. The category II methods significantly outperform the category I methods. It indicates that the combination of two measures will improve the performance. The measures in different visual areas are not with the complete duplicates functions in visual evaluations. Furthermore, the category III method, HVQA outperforms any other combinations. That is, the proposed metric is the most efficient for perceptual video quality evaluation.

Table 2 shows the PCC of state-of-the-art metrics perform on four kinds of distortions in the LIVE database. Fig. 2 shows the performance comparison between these metrics as bar charts. It shows that the proposed metric performs the best on three kinds of distortions (Wireless, IP, and H.264). For the MPEG2 distortions, the proposed HVQA metric is also competitive with the VQM metric, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

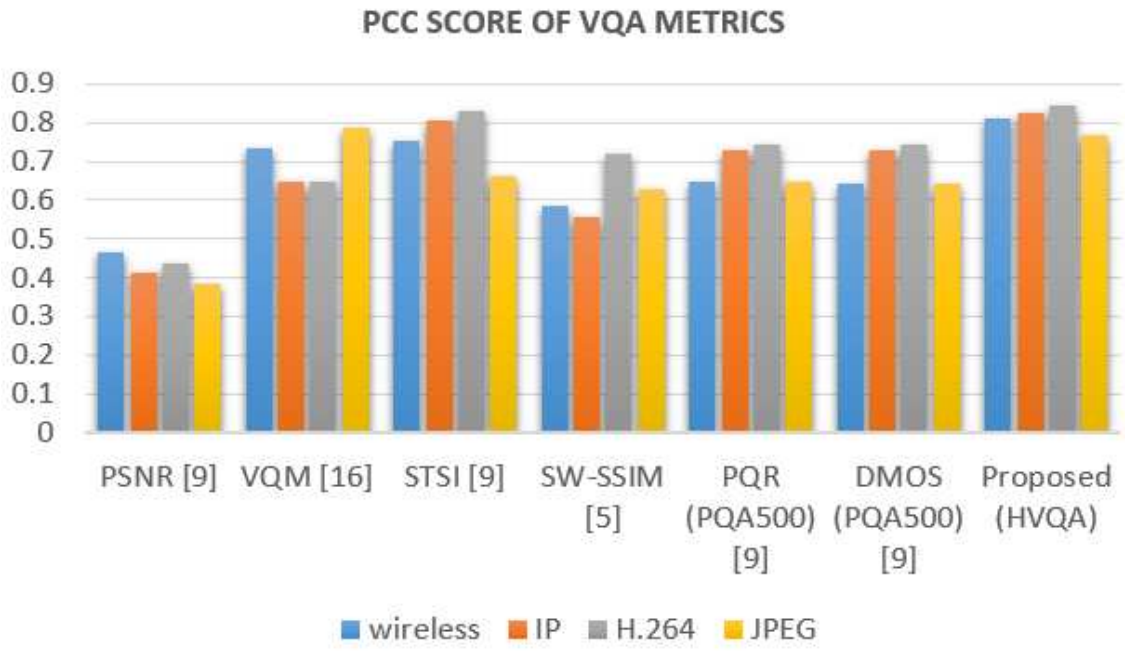

Figure 2. PCC SCORES OF VQA METRICS ON EACH KIND OF DISTORTION IN LIVE DATABASE.

performs the best for this type of distortions. Thus, the HVQA metric is rather robust to various types of the video distortions. It is observed that the HVQA metric significantly outperforms the STSI index for all the four types of the distortions. This coincides with the former analysis that, the similarities of the area V4 and visual attention can improve the performance.

Fig. 3 shows the scatter plot of the DMOS against the objective computational score. It is observed that the proposed HVQA metric performs well on videos from the low qualities to the high qualities.

3. Conclusion

In this letter, a hierarchical VQA metric has been proposed inspired by the primate visual cortex. The neuronal processing of the visual information in the sequential visual areas are modeled with the corresponding measures. Experimental results show that the proposed metric significantly outperforms the state-of-the-art VQA metrics.

References

1. M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, and K. Brunnstrom, T. Subjective multimedia quality assessment. *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences* **2006**, *vol. E89-A, no. 11*, pp. 2920-2932, Nov 2006.
2. T. Yamada, Y. Miyamoto, M. Serizawa, and T. Nishitani, T. Reduced- reference video quality estimation using representative luminance. *IEICE Trans. Fundamentals of Electronics, Communications and Com- puter Sciences* **2006**, *vol. E95-A, no. 5*, pp. 961-986, May 2012.
3. Z. Wang and A. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine* **2009**, *vol. 26, no. 1*, pp. 98-117, Jan. 2009.
4. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *EEE Trans. Image Process.* **2004**, *vol. 13, no. 4*, pp. 600-612, Apr. 2004.
5. Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *J. Opt. Soc. Amer. A.* **2007**, *vol. 24, no. 12*, pp. B61-B69, Apr. 2007.
6. A. Moorthy and A. Bovik. Efficient video quality assessment along temporal trajectories. *IEEE Trans. Circuits and Syst. Video Technol.* **2010**, *vol. 20, no. 11*, pp. 1653-1658, Nov. 2010.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2017 doi:10.20944/preprints201704.0088.v1

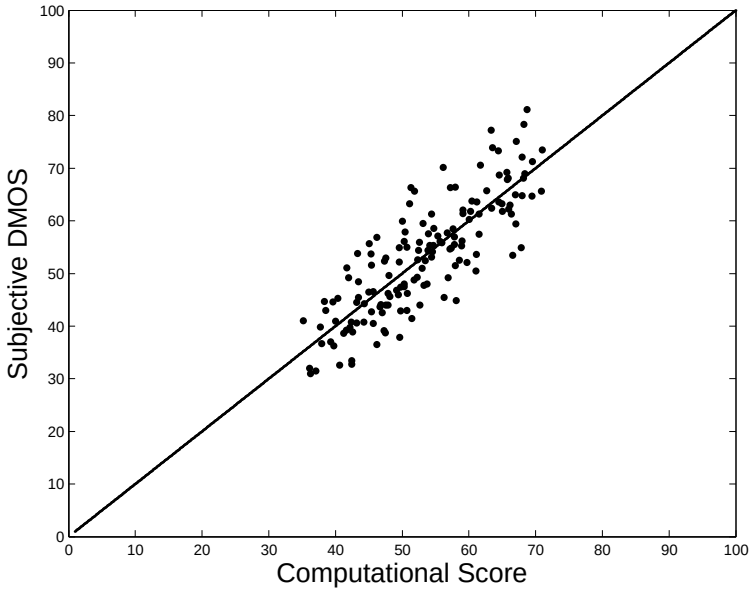

Figure 3. Scatter plot of DMOSs against scores predicted by HVQA.

7. J. Wu, W. Lin, G. Shi, and A. Liu. Perceptual quality metric with internal generative mechanism. *IEEE Trans. Image Process.* **2013**, *vol. 22, no. 1*, pp. 43-54, Jan. 2013.
8. X. Zhang, X. Feng, W. Wang, and W. Xue. Edge strength similarity for image quality assessment. *IEEE Signal Process. Lett.* **2013**, *vol. 22, no. 1*, pp. 319-322, Apr. 2013.
9. Y. Wang, T. Jiang, S. Ma, and W. Gao. Novel spatio-temporal structural information based video quality metric. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *vol. 22, no. 7*, pp. 989-998, July 2012.
10. N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Trans. Pattern Analysis and Machine Intelligence.* **2013**, *vol. 35, no. 8*, pp. 1847-1871, 2013.
11. D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex.* **2013**, *vol. 1, no. 1*, pp. 1-47, 1991.
12. K. Dabov, A. Foi, and K. Egiazarian. Video denoising by sparse 3d transform-domain collaborative filtering. *in Proc. European Signal Process. Conf., EUSIPCO 2007.* Poznan, Poland; Sep. 2007.
13. Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters.* **2008**, *vol. 44, no. 13* pp. 800-801, 2008.
14. H. Li and K. N. Ngan. Saliency model based face segmentation in head-and-shoulder video sequences. *J. Visual Communication and Image Representation, Elsevier Science.* **2008**, *vol. 19, no. 5* pp. 320-333, 2008.
15. H. Li and K. N. Ngan. A co-saliency model of image pairs. *IEEE Trans. Image Pro- cess.* **2011**, *vol. 20, no. 12* pp. 3365-3375, 2011.
16. M. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Trans. Broadcasting.* **2004**, *vol. 50, no. 3* pp. 312-322, 2004.
17. K. Seshadrinathan and A. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. Image Process.* **2004**, *vol. 19, no. 2* pp. 335-350, Feb. 2010.
18. K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack. Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.* **2010**, *vol. 19, no. 6* pp. 1427-1441, Jun. 2010.