Video quality assessment using motion-compensated temporal filtering and manifold feature similarity

Yang Song*, Mei Yu, Gangyi Jiang*, Feng Shao, Zongju Peng
Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, Zhejiang, China

* songyang_ac@126.com (YS); jianggangyi@126.com (GJ)

Abstract

Well-performed Video quality assessment (VQA) method should be consistent with human visual systems for better prediction accuracy. In this paper, we propose a VQA method using motion-compensated temporal filtering (MCTF) and manifold feature similarity. To be more specific, a group of frames (GoF) is first decomposed into a temporal high-pass component (HPC) and a temporal low-pass component (LPC) by MCTF. Following this, manifold feature learning (MFL) and phase congruency (PC) are used to predict the quality of temporal LPC and temporal HPC respectively. The quality measures of the LPC and the HPC are then combined as GoF quality. A temporal pooling strategy is subsequently used to integrate GoF qualities into an overall video quality. The proposed VQA method appropriately processes temporal information in video by MCTF and temporal pooling strategy, and simulate human visual perception by MFL. Experiments on publicly available video quality database showed that in comparison with several state-of-the-art VQA methods, the proposed VQA method achieves better consistency with subjective video quality and can predict video quality more accurately.

Introduction

The rapidly growing popularity of such digital consumer electronic devices as smartphones and portable computers has rendered video applications ubiquitous in our daily lives. Prior to being received by the users, video information needs to pass through several stages in communication systems, and is inevitably affected by noise and various kinds of distortion. Therefore, an accurate video quality assessment (VQA) method is needed to improve system performance and the quality of the users’ viewing experience.

Videos can be considered as orderly arrangements of several images, called frames. Therefore, a video contains both intra-frame spatial information and inter-frame temporal information. An effective VQA method should thus take both two aspects into consideration. In the last decade, the processing of spatial information in images has drawn an increasing amount of research interest. Due to a better understanding of the human visual system (HVS) [1–2] and advances in natural scene statistics (NSS) [3], a series of image quality assessment (IQA)
methods have been proposed [4–5]. In the early stages of video quality evaluation, traditional IQA methods were used to predict the quality of each frame, and then computed the average of all frames’ qualities as an overall video quality [6]. The image quality of each frame clearly contributes considerably to overall video quality. However, such methods overlook the importance of temporal information, which limits their effectiveness. To overcome this disadvantage, several researchers have attempted to integrate temporal information into their methods. The relevant methods mainly used global motion to represent temporal information in video. Seshadrinathan et al. [7] considered motion information as a video feature and proposed motion-tuned, spatio-temporal quality assessment of natural video (MOVIE). To further investigate the HVS response to motion information in videos, Li et al. explored the effects of the spatio-temporal contrast sensitivity function. Meanwhile, by analyzing the characteristics of distortion in videos, a noise decoupling-based VQA method has been proposed [8]. Zhang et al. [9] exploited the visual masking effect to process the human perception of distortion in videos, and proposed a perception-based VQA method. In recent years, researchers have also attempted to simultaneously process spatial and temporal information by three-dimensional (3D) decomposition. In [10], Torkamani-Azar considered videos as 3D matrices, and used 3D singular value decomposition (3D-SVD) to extract 3D singular vectors as video features, and this 3D-SVD-based VQA method works well to evaluate video quality.

From the perspective of neuro-biology, the ultimate goal of VQA is to simulate the response of human visual systems. Previous studies have already revealed that manifold is fundamental to perception [11]. Given visual information like videos, which can be considered as a set of high-dimensional data, manifold learning aims at discovering the geometric properties existing inside the data. Therefore, it can be utilized to eliminate the redundancies in videos and extract the essence structure as video features. In recent years, several manifold learning methods have been proposed [12–14]. These methods have been widely used in several image/video processing fields, such as face recognition [15], image classification [16], etc. However, there is still less work focusing on applying manifold learning methods on predicting visual quality, especially for video quality.

According to above discussion, it is evident that the two most challenging issues in VQA are temporal description and simulation of human perception. Specifically, in this paper, we analyze wavelet coding theory [17] and consider its method of temporal decomposition to be a reference. We introduce motion-compensated view filtering (MCTF) from wavelet coding to decompose videos in the temporal domain. Then, in order to simulate human visual perception, Orthogonal Locality Preserving Projection (OLPP) algorithm [18] is employed to extract manifold feature. Finally, an asymmetric temporal pooling strategy is adopted to obtain an overall video quality. This newly proposed VQA method takes the following unique features:

1. According to the contracture of video, we utilize MCTF to decompose video into different frequency components;
2. By analyzing the characteristics of different frequency components, we deploy appropriate method to evaluate each component’s quality and integrate both qualities into video quality;
3. To ensure the VQA method incorporate with human visual characteristics, we use manifold learning as a perceptual approach to extract features.

The rest of this paper is organized as follows: Section 2 introduces each part of the proposed VQA method in detail. Experiments conducted on the Laboratory for Image & Video Engineering (LIVE) video quality database are described in Section 3. Directions for further research in the area are discussed in Section 4.
Materials and methods

To address the difficulty in representing temporal information, we deal with it at both group of frame (GoF) level and video level. At the GoF level, MCTF is used to decompose temporal information into two different parts, namely, temporal high-pass component (HPC) and temporal low-pass component (LPC), whereas at the video level, a temporal pooling strategy is adopted. In order to accurately predict the qualities of both the HPC and the LPC, we use manifold learning and phase congruency (PC) similarity to simulate human visual perception. Based on this analysis, we propose a video quality assessment method using MCTF and manifold feature similarity. Fig 1 shows the framework of the proposed VQA method. It consists of five sequential processing modules. GoFs are first decomposed into a temporal HPC and a temporal LPC by MCTF. The quality of each temporal HPC and temporal LPC is then separately assessed, following which they are integrated as GoF quality. Finally, an overall video quality is obtained by the temporal pooling strategy of all GoF qualities.

Temporal filtering in GoFs

It is well known that different frequency components consist varying information in an image, and distortion appears in different frequency components unevenly degrades image quality. This effect also exists in video. Specifically, the lower frequency component consists of still objects and structural information in the video, whereas the higher frequency component represents detail information concerning moving objects. As a result, it is necessary to decompose the video into different frequency components and appropriately predict the quality of each.

Traditional temporal filters directly decompose pixels at the same location in several frames. Because temporal motion widely exists in videos, which leads to scene displacement in adjacent frames, traditional temporal filters cannot thoroughly decompose video. In proposed VQA method, we use MCTF to implement temporal filtering in GoFs. The MCTF can decompose GoFs along the trajectory of the motion objects and archive better decomposition performance. In implementing MCTF on a GoF, two adjacent frames are filtered firstly. The filtering procedure can then be divided into two steps: motion compensation (MC) and temporal filtering (TF). In the MC step, let $l_{n+1}$ and $l_n$ denote two adjacent frames in a GoF. We first take $l_n$ as a reference frame and search the matching block in $l_{n+1}$ using a three-step search algorithm [19] to obtain motion vector $mv_{n+1→n}$. Then, the mapping from $l_n$ to $l_{n+1}$, denoted by $M_{n+1→n}$ can be acquired by using the motion vectors of both vertical and horizontal directions.

![Fig 1. Framework for the proposed VQA method.](https://doi.org/10.1371/journal.pone.0175798.g001)
Eventually, \( M_{n+1} \) can be subsequently used to transform \( l_{n+1} \) to motion-compensated frame \( MC_{n+1} \) using

\[
MC_{n+1}[x + mv_{n+1}^H(x,y), y + mv_{n+1}^V(x,y)] = l_{n+1}(x,y)
\]

(1)

where \( mv_{n+1}^H(x,y) \) is the horizontal motion vector from \( l_{n+1} \) to \( l_n \) in \((x,y)\), and \( mv_{n+1}^V(x,y) \) is the horizontal motion vector from \( l_{n+1} \) to \( l_n \) in \((x,y)\).

In the TF step, we use a lifting-based technique to decompose \( l_n \) and \( MC_{n+1} \). The lifting technique is an efficient implementation of the wavelet transform that uses low memory and is not computationally complex. Let \( H \) denote temporal HPC and \( L \) denote temporal LPC. The detailed implementation of decomposing \( l_{n+1} \) and \( MC_{n+1} \) can be represented by Eq (2):

\[
\begin{align*}
H(x,y) &= \frac{1}{\sqrt{2}} \left[ l_n(x,y) - MC_{n+1}(x,y) \right] \\
H(x,y) &= \frac{1}{\sqrt{2}} \left[ l_n(x,y) + \frac{1}{2} [MC_{n+1}(x,y) - H(x,y)] \right]
\end{align*}
\]

(2)

where, \((x,y)\) represents pixel location.

Fig 2 shows the implementation of MCTF in a four-frame GoF. Following the decomposition of adjacent frames (F1 and F2, F3 and F4), two temporal HPC frames (H1, H2) and two temporal LPC frames (L1, L2) are obtained. Let \( CHPC \) GoF and \( CLPC \) GoF denote the temporal HPC and the temporal LPC of the GoF, respectively. Then, \( CHPC \) GoF can be derived by MCTF in \( H_1 \) and \( H_2 \), and \( CLPC \) GoF is obtained by deploying MCTF to \( L_1 \) and \( L_2 \).

Note that distortion, such as blurring or blockiness, can change the result of block-based motion estimation, which affects the results of motion vector searching and causes the performance degradation of MCTF. Consequently, when a distorted GoF is processed by MCTF, we use the motion vector obtained from corresponding reference GoF instead. Fig 3 shows the result of implementing MCTF in a four-frame GoF, which is a randomly picked GoF in a video called “Pedestrian Area” from the LIVE video database. According to Fig 3(B) and 3(C), the temporal HPC consists of detail information regarding moving objects, whereas the temporal LPC contains the structural information of original scenes in the GoF and reserves all still objects.

**Temporal LPC quality metric**

Natural scene videos are highly structured and can be seen as a high-dimensional set of data. Therefore, manifold learning can be applied in video to reduce the data’s dimension and
extract the low-dimensional features, which can accurately reflect the intrinsic property of the video. From the results of the MCTF in Fig 3(B), we can find out that temporal LPC contains the most content and scene in the original GoF, and perceives to be quite similar to the original. Thus, it is reasonable to conclude that temporal LPC retains the most essential characteristics of original videos. Based on above analysis, we use manifold learning in the proposed VQA method to extract the distorted features of temporal LPC.

**Feature extraction matrix learning.** To extract the manifold feature from temporal LPC, firstly, 10,000 overlapped image patches with a size of 8×8 are randomly picked to build a training set. Next, the OLPP algorithm are employed to obtain the projection matrix in training set. Finally, the projection matrix, which is utilized as feature extraction matrix, are used to extract the manifold features. The specific implementation of feature extraction matrix learning is stated as follows.

Before using OLLP to train the feature extraction matrix, we first apply principle component analysis (PCA) to reduce the dimension of input sample $Y$ and only retain first 8 principle components of $Y$ for training (the detail implementation of PCA can be referred in [20]). Meanwhile, [21] indicated that whitening can be used to simulate the working mechanism that Lateral Geniculate Nucleus (LGN) process visual information. Therefore, $Y$ can be whitened into $Y^w$ by

$$Y^w = Y \times W$$  \hspace{1cm} (3)

where $W$ is the whitened matrix, it can be calculated by eigenvalue and covariance matrix of $Y$.

Let $G$ denote a graph with $m$ nodes, and the $n$th node represents the whitened sample $yw n$. Two nodes can be linked when they are adjacent, i.e., $yw a$ is among the nearest neighbors of $yw b$. Moreover, if node $a$ and node $b$ are linked, the weight $S_{ab}$ can be set as $e^{-||yw a - yw b||^2}$, otherwise $S_{ab} = 0$. In order to model the local manifold structure, we define $S$ as weight matrix.

Then, the diagonal matrix $\Phi$ can be acquired by $\Phi_{aa} = \sum_{b=1}^{N} S_{ab}$, and the Laplacian matrix $L$ is calculated as $L = \Phi - S$. Let $\{p_1, \ldots, p_n\}$ denote the orthogonal basis vector, it can be
defined as

\[
\begin{align*}
\mathbf{p}_n &= V_{\lambda_{\text{min}}} \left[ (\mathbf{Y}^T \Phi \mathbf{Y}^T)^{-1} \mathbf{Y}^T \mathbf{L} \mathbf{Y}^T \right], \quad n = 1 \\
\mathbf{p}_n &= V_{\lambda_{\text{min}}} \left\{ \mathbf{I} - (\mathbf{Y}^T \Phi \mathbf{Y}^T)^{-1} \mathbf{P}^{(n-1)} [\mathbf{Q}^{(n-1)}]^{-1} [\mathbf{P}^{(n-1)}]^T \right\} \\
&\quad \cdot (\mathbf{Y}^T \Phi \mathbf{Y}^T)^{-1} \mathbf{Y}^T \mathbf{L} \mathbf{Y}^T, \quad \text{others}
\end{align*}
\] (4)

where \( V_{\lambda_{\text{min}}} \) represents the eigenvector corresponding to the smallest non-zero eigenvalue, and the orthogonal basis function, denoted by \( \mathbf{F}_Q \), is expressed as

\[
\mathbf{F}_Q^{(n-1)} = [\mathbf{P}^{(n-1)}]^T (\mathbf{Y}^T \Phi \mathbf{Y}^T)^{-1} \mathbf{P}^{(n-1)}
\] (5)

Let \( M = \{ \mathbf{p}_1, \ldots, \mathbf{p}_n \} \) denote the transformation matrix. According to the PCA result, \( r \) is set to 8. Finally, the transformation matrix should be transformed from whitened space to original one as Eq (6) illustrated,

\[
\mathbf{M}_{\text{opt}} = W \times M
\] (6)

where \( \mathbf{M}_{\text{opt}} \) is the optimal projection matrix, which can be used to extract image’s manifold feature.

It should be noted that in order for the optimal projection to more accurately reflect the essential features of temporal LPC, we used the temporal LPC to construct the training set. Specifically, we randomly selected 10 GoFs from reference videos in the LIVE video quality database and extracted 10,000 blocks from their temporal LPCs as the training set. Fig 4 shows the training set selected for the proposed VQA method.

**Manifold feature similarity.** The previously obtained 8×8 optimal projection matrix \( \mathbf{M}_{\text{opt}} \) can then be used to extract the manifold features. Let \( \text{CLPC ref}(i) \) and \( \text{CLPC dis}(i) \) denote blocks in the reference and the distorted temporal LPC, and \( \text{MF}_{\text{ref}}(i) \) and \( \text{MF}_{\text{dis}}(i) \) denote the manifold features of the reference and the distorted temporal LPC, which can be obtained by

\[
\text{Fig 4. Training set from randomly selected temporal LPCs.}
\]

https://doi.org/10.1371/journal.pone.0175798.g004
Eq (7) and Eq (8):

\[ MF^\text{ref}_i(k_1, k_2) = \sum_{k_1=1}^{8} \sum_{k_2=1}^{8} C_{\text{ref}}^\text{LPC}(k_1, k_2) \cdot M_{\text{opt}}(8 - k_1, 8 - k_2) \quad (7) \]

\[ MF^\text{dis}_i(k_1, k_2) = \sum_{k_1=1}^{8} \sum_{k_2=1}^{8} C_{\text{dis}}^\text{LPC}(k_1, k_2) \cdot M_{\text{opt}}(8 - k_1, 8 - k_2) \quad (8) \]

After obtaining the manifold features, the next step is to calculate the image block qualities by using these features. According to the similarity measurement defined in SSIM (Structural Similarity), the manifold feature similarity for each block in temporal LPC can be acquired as Eq (9). This is a commonly used method to calculate the similarity between two positive number sets. The result of Eq (9) is in the range of 0 to 1, and 1 implies a perfect match between two set of numbers.

\[ q_{\text{block}}^\text{LPC} = \frac{2 \times MF_{\text{ref}}(i) \times MF_{\text{dis}}(i) + C_1}{MF_{\text{ref}}^2(i) + MF_{\text{dis}}^2(i) + C_1} \quad (9) \]

where \( C_1 \) is a small constant to ensure that the denominator is non-zero,

Then, the quality of temporal LPC can be integrated by averaging all blocks’ qualities. Let \( q^\text{LPC} \) denote the quality of temporal LPC. It can be obtained as

\[ q^\text{LPC} = \frac{1}{k} \sum_{i=1}^{k} q_{\text{block}}^\text{LPC} \quad (10) \]

where \( k \) is the total number of blocks in an LPC.

**Temporal HPC quality metric**

As previously mentioned in Section 2, the temporal HPC of a GoF contains information regarding moving objects and related details. Phase is an important image feature that captures a considerable amount of detail information of an image. As a result, phase feature can be used to evaluate the quality of temporal HPC. Previous researches have revealed that the HVS is highly sensitive to pixels with high phase congruency (PC) [22]. Thus, in the proposed VQA method, we extract the PC from temporal HPC as distortion features.

It is well-known that the visual cortex can be satisfactorily simulated by the Log-Gabor filter. In this paper, we use the responses of the Log-Gabor filter to calculate the PC of the temporal HPC. Specifically, Eq (11) is adopted to calculate PC. The detail explanation of Eq (11) can be found in [22].

\[ PC(m) = \frac{E_s(m)}{\varepsilon + \sum_s A_s(m)} \quad (11) \]

where \( A_{s,o}(m) \) represents local amplitude of response in Log-gabor filter, and \( E_s(m) \) denotes local energy. \( s \) and \( o \) is defined as scale and orientation of the filter. \( \varepsilon \) is a small constant to avoid denominator being zero.

Following the calculation of the PC features as above, we predict the quality of the temporal HPC using these features. Let \( PCHPC\text{ref} \) and \( PCHPC\text{dis} \) denote the PC features of the reference temporal HPC and the distorted temporal HPC, respectively. The similarity measurement method used in manifold feature similarity calculation is then used to obtain the quality
of the distorted temporal HPC, as illustrated in Eq (12):

\[ q_{\text{HPC}} = \frac{2 \times PC_{\text{HPC}} \times PC_{\text{HPC}}^* + C_2}{(PC_{\text{HPC}}^*)^2 + (PC_{\text{HPC}})^2 + C_2} \]  

(12)

where \( C_2 \) is a small positive constant to ensure a non-zero denominator.

**GoF quality pooling**

The quality of \( n \)th GoF, denoted by \( q_{\text{GoF}} n \), can be obtained by a combination of the quality of the temporal HPC \( q_{\text{HPC}} n \) and that of the temporal LPC \( q_{\text{LPC}} n \). In the proposed VQA method, we use a linear weighting summation model to calculate \( q_{\text{GoF}} n \) as Eq (13) illustrated. Because video processing contains huge data volume, we adopt linear summation instead of more sophisticated regression model for its low computational complexity.

\[ q_{\text{GoF}} n = \omega_1 \cdot q_{\text{HPC}} n + \omega_2 \cdot q_{\text{LPC}} n \]  

(13)

where \( \omega_1 \) and \( \omega_2 \) are the weights assigned to temporal HPC and temporal LPC. It is expected that the quality of temporal LPC has larger impact on overall GoF quality and so \( \omega_1 < \omega_2 \). In our method, \( \omega_1 \) is set to 0.3 through performance tuning, and \( \omega_2 \) is set to 0.7. Details of these determinations is illustrated in Section 3.2.

**Temporal pooling**

Following the calculation of the qualities of all GoFs, a VQA method needs to integrate all GoF qualities into an overall video quality. However, simply averaging all GoF qualities does not consist with human perception, and is likely to degrade the performance of prediction. It is thus necessary to simulate several HVS characteristic while combining GoF qualities into video quality. Because observers are more sensitive to degradation in video quality than improvement, we adopt the implementation presented in [23] to simulate such asymmetric responses to fluctuation in GoF qualities. We first adjust GoF quality \( q_{\text{GoF}} n \) to intermediate GoF quality \( q_{\text{GoF'}} n \) according to Eq (14):

\[ q_{n}^{\text{GoF'}} = \begin{cases} 
q_{n-1}^{\text{GoF'}} + a^+ q_{n}^{\text{GoF}}, & q_{n}^{\text{GoF}}>0 \\
q_{n-1}^{\text{GoF'}} + a^- q_{n}^{\text{GoF}}, & q_{n}^{\text{GoF}} \leq 0
\end{cases} \]  

(14)

where \( a^+ \) and \( a^- \) embody asymmetric behavior, and

\[ q_{\text{GoF}} n = q_{n}^{\text{GoF'}} - q_{n-1}^{\text{GoF'}} \]  

(15)

In the proposed VQA method, \( a^+ \) is set to 0.09 and \( a^- \) to 0.8 through performance tuning. Finally, the overall video quality \( Q \) is calculated by averaging all the intermediate GoF quality.

\[ Q = \frac{1}{N} \sum_{n=1}^{N} q_{n}^{\text{GoF'}} \]  

(16)

where \( N \) is the number of GoFs in a video.
Results and discussion

Subjective database and performance index

The LIVE video quality [24] database is used to evaluate the performance of the proposed VQA method. The LIVE video quality database consists of 10 reference videos and 150 distorted videos generated from the reference videos. The distorted videos are created by using four types of distortion: Wireless network transmission Distortion (WD), IP network transmission (IP) distortion (IP), H.264 compression distortion (H264), and MPEG-2 compression distortion (MPEG-2). All reference videos and distorted videos have a resolution of 768×432 and the frame rate ranged from 25 fps to 50 fps.

Three measures are employed as performance indexes to evaluate the performance of the proposed VQA method: the Pearson linear correlation coefficient (PLCC), Spearman’s rank-order correlation coefficient (SROCC), and root-mean-square error (RMSE). Detailed formulation regarding these can be reviewed in [25]. In general, a higher PLCC represents better correlation between predicted quality and subjective assessment of quality. The SROCC measures the monotonicity of predicted quality, whereas the RMSE measures error in predicted quality. A smaller RMSE indicates better prediction performance.

Parameterization

Two sets of parameters need to be determined in the proposed VQA method, i.e., \( \omega_1, \omega_2 \) and \( a^+, a^- \). In tuning \( \omega_1 \) and \( \omega_2 \), both \( a^+ \) and \( a^- \) are set to 1. \( \omega_1 \) is changed from 0 to 1 in increments of 0.1. Fig 5(A) shows the result of tuning of \( \omega_1 \). When \( \omega_1 \) is set to 0.3, the highest value of the PLCC is obtained. Since \( \omega_1 + \omega_2 = 1 \), \( \omega_2 \) is set to 0.7. The tuning result shows that the quality of the temporal LPC contributes most to overall quality. It further confirms that the temporal LPC contained the greatest number of items of information from the original video. Once \( \omega_1 \) and \( \omega_2 \) are fixed, \( a^+ \) and \( a^- \) are set by the same method. \( a^+ \) is changed from 0 to 0.4 in increments of 0.01, and \( a^- \) is changed from 0 to 4 in increments of 0.1. The results can be seen in Fig 5(B). When \( a^+ = 0.09, a^- = 0.8 \), the PLCC reaches its peak. The significant difference in value confirm the hypothesis of the asymmetry of human responses to fluctuations in quality: to wit, observers are more sensitive to quality degradation than quality improvement while watching videos.

Determination of GoF size

In proposed method, the MCTF is implemented on GoFs. Therefore, it is necessary to explore whether the GoF size will have effect on final performance. As the requirement in
implementation of lifting-based wavelet transform, the GoF number should be set as \( n \)-th power of 2. Consequently, Table 1 is listed to illustrate the effect of GoF size.

By observing the experimental results in Table 2, it can be conclude that the performances are nearly equivalent while the GoF numbers are 4, 8, or 16, and when the GoF number is greater than 32, the performance degraded sharply. Furthermore, in implementation, the larger GoF number will inevitably introduce tremendous computational complexity. Therefore, in proposed VQA method, the final GoF number is set as 4.

Impact of each module in the proposed VQA method

To verify the impact of each module employed in the proposed VQA method, we design another three plans that partly choose several modules in proposed method for comparison, denoted by Plan-A, Plan-B and Plan-C, respectively. In Plan-A, we use natural image instead of randomly picked temporal LPC to build training set, and other implementations are same with proposed method. In Plan-B, motion compensation is not included in temporal decomposition, and other implementations are same with the proposed method. In Plan-C, all GoF qualities are averaged to yield the overall video quality and other implementation are same with proposed method. By comparing Plan-A with proposed method, we can conclude that the optimal projection matrix trained by a proper training set coincides to a greater extent with human perception. The performance improvement between Plan 2 and the proposed method shows that motion compensation profits the temporal filtering in the temporal decomposition of GoFs. Finally, by using the temporal pooling strategy instead of simply averaging the quality measures of all GoFs, the proposed method outperforms Plan 3. In summary, each module employed in the proposed VQA method plays a positive role in performance improvement.

Overall prediction performance

Table 3 lists the performance indexes of the proposed VQA method in the LIVE video quality database. For comparison, we also provide the results of several methods in Table 4. These

| GoF number | PLCC  | SROCC | RMSE   |
|-----------|-------|-------|--------|
| 4         | 0.8472| 0.8298| 4.9764 |
| 8         | 0.8471| 0.8300| 4.9788 |
| 16        | 0.8464| 0.8271| 4.9889 |
| 32        | 0.8217| 0.8003| 5.1324 |
| 64        | 0.8035| 0.7961| 5.4231 |

Table 2. PLCC comparison for each module of the proposed VQA method.

| Details       | WD    | IP    | H264   | MPEG   | ALL    |
|---------------|-------|-------|--------|--------|--------|
| Plan-A        | MC+TP | 0.8406| 0.7711 | 0.7283 | 0.8384 | 0.8353 |
| Plan-B        | PTS+TP| 0.8255| 0.7671 | 0.7462 | 0.8136 | 0.8107 |
| Plan-C        | MC+PTS| 0.8280| 0.7913 | 0.7506 | 0.8247 | 0.8235 |
| Proposed method| MC+PTS+TP| 0.8591| 0.7773| 0.7385| 0.8570| 0.8472|

MC: Motion compensation in temporal filtering. PTS: Proper training set (randomly selected temporal LPC) in MFL. TP: using temporal pooling strategy to combine GoF quality measures.
comparative methods include the traditional image quality metric with temporal averaging (PSNR, SSIM), methods using motion information (MOVIE, the method proposed in [8]), method simulating the HVS working mechanism (the metric proposed in [9]), and method adopting 3D transformation (VRF) and the method standardized by Video quality expert group (VQEG) [26]. The best performance is shown in bold in the table. We see from Table 4 that the proposed methods achieve the best indexes for distortion like WD, IP and MPEG-2. As for H264 distortion, the indexes of proposed VQA method are not as accurate as other VQA methods. Essentially, the H.264 compression will introduce blocking effect in videos. Meanwhile, PC feature is much more sensitive to those artificial edges caused by blocking effect than human eyes. Therefore, slight quality degradation caused by block effect can be exaggerated by PC-based quality metric adopted in proposed method. As a result, the prediction accuracy of proposed method will be declined for H.264 compressed video. However, the proposed VQA method outperformed all other VQA methods in terms of overall performance for all distorted videos (ALL) in the LIVE video quality database. Taking all indicators into consideration, the proposed VQA methods yield the highest correlation with subjective quality and can predict video quality more accurately.

We also show the scatter plots of the proposed VQA method in Fig 6. The horizontal axis denotes the predicted qualities obtained by proposed method and the vertical axis denotes the subjective qualities provided by LIVE database. These scatter plots reflect the approximate linear correlation between the prediction qualities and subjective qualities. Fig 6 shows that the predictive quality of the proposed VQA method was highly correlated with subjective assessments of quality.

**Conclusions**

In this paper, we propose a video quality metric using motion-compensated temporal filtering (MCTF) and manifold feature similarity. The main idea underlying this method is to decompose videos in the temporal domain and appropriately predict the qualities of temporal LPC

### Table 3. Performance on separate types of distortion.

|        | PLCC   | SROCC  | OR   | RMSE  |
|--------|--------|--------|------|-------|
| WD     | 0.8591 | 0.8359 | 0    | 4.8332|
| IP     | 0.7573 | 0.7513 | 0    | 6.4879|
| H264   | 0.7385 | 0.7230 | 0.025| 7.6598|
| MPEG-2 | 0.8570 | 0.8414 | 0    | 4.9124|
| ALL    | 0.8472 | 0.8298 | 0    | 4.9764|

### Table 4. Performance comparison on the LIVE VQA database.

|        | WD  | IP  | H264 | MPEG2 | ALL  |
|--------|-----|-----|------|-------|------|
| PSNR   | 0.4675 | 0.4108 | 0.4385 | 0.3856  | 0.4035  |
| SSIM   | 0.5401 | 0.5119 | 0.6656 | 0.5491  | 0.5444  |
| VQM    | 0.7325 | 0.648  | 0.6459 | 0.786   | 0.7236  |
| MOVIE  | 0.8386 | 0.6772 | 0.7902 | 0.7595  | 0.8116  |
| STAQ   | 0.5684 | 0.7080 | **0.8778** | 0.7988  | 0.7192  |
| VRF    | 0.7708 | 0.7453 | 0.7062 | 0.6019  | 0.6983  |
| Metric in [8] | —   | —    | —    | —     | **0.8460**  |
| Metric in [9] | 0.8280 | 0.7010 | 0.8160 | 0.7420  | 0.8150  |
| Proposed* | **0.8591** | **0.7573** | 0.7385 | **0.8570**  | **0.8472**  |
and temporal HPC generated from temporal decomposition. Specifically, we use MCTF to decompose a GoF into different frequency components. According to the characteristics of both the frequency components and human perception characteristics, we extract manifold features and the phase congruency in temporal LPC and temporal HPC, respectively, and then calculated feature similarity as GoF quality. Finally, a temporal pooling strategy is used to obtain an overall video quality. Experiments on the LIVE video quality database shows that the proposed VQA method performed satisfactorily in predicting video quality. In future work, some outstanding issues need to be considered, such as a better temporal pooling strategy as well as a temporal decomposition method.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grants 61271270, 61271021, 61311140262 and U1301257, the National High-tech R&D Program of China (2015AA015901), Zhejiang Provincial Natural Science Foundation of China (LY15F010005, Y16F010010) and the K. C. Wong Magna Fund at Ningbo University.

Author Contributions

Conceptualization: YS MY GJ.
Formal analysis: YS MY GJ.
Funding acquisition: MY GJ.
Investigation: YS MY GJ FS ZP.
Methodology: YS MY GJ.

Project administration: MY GJ.

Resources: YS MY GJ FS ZP.

Software: YS.

Supervision: MY GJ.

Validation: YS.

Visualization: YS.

Writing – original draft: YS.

Writing – review & editing: YS MY GJ FS ZP.

References

1. Park MC, Mun S. Overview of Measurement Methods for Factors Affecting the Human Visual System in 3D Displays. Journal of Display Technology. 2015; 11: 877–888.

2. Omar H, Hammett ST. Perceptual biases are inconsistent with Bayesian encoding of speed in the human visual system. Journal of Vision. 2015; 15(2): 9. https://doi.org/10.1167/15.2.9 PMID: 25761348

3. Song L, Chen C, Xu Y, Xue G. Blind image quality assessment based on a new feature of nature scene statistics. 2014 IEEE Visual Communications and Image Processing Conference. 2014:37–40.

4. Shao F, Li K, Lin W, Jiang G, Yu M. Using Binocular Feature Combination for Blind Quality Assessment of Stereoscopic Images. IEEE Signal Process Lett. 2015; 22: 1548–1551.

5. Zhang L, Zhang L, Bovik AC. A Feature-Enriched Completely Blind Image Quality Evaluator. IEEE Trans Image Process. 2015; 24: 2579–2591. https://doi.org/10.1109/TIP.2015.2426416 PMID: 25915960

6. You J, Korhonen J, Perkis A. Spatial and temporal pooling of image quality metrics for perceptual video quality assessment on packet loss streams. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2010: 1002–1005.

7. Seshadrinathan K, Bovik AC. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. IEEE Trans Image Process. 2010; 19: 335–350. https://doi.org/10.1109/TIP.2009.2034992 PMID: 19846374

8. Li S, Ma L, Ngan KN. Full-Reference Video Quality Assessment by Decoupling Detail Losses and Additive Impairments. IEEE Trans Circuits Syst Video Technol. 2012; 22:1100–1112.

9. Zhang F, Bull D. A Perception-based Hybrid Model for Video Quality Assessment. IEEE Trans Circuits Syst Video Technol. 2015:1–12.

10. Torkamani-Azar F, Imani H, Fathollahian H. Video quality measurement based on 3-D. Singular value decomposition. J Vis Commun Image Represent. 2015; 27:1–6.

11. Seung HS, Dd L. Cognition: The manifold ways of perception. Science. 2000; 290: 2268–2269. PMID: 11188725

12. He X. Locality preserving projections. Adv Neural Inf Process Syst. 2005; 45: 186–197.

13. Mukund B, Schwartz EL. The isomap algorithm and topological stability. Science. 2002; 295: 7–7. https://doi.org/10.1126/science.295.5552.7a PMID: 11778013

14. Silva VD, Tenenbaum JB. Global versus Local Methods in Nonlinear Dimensionality Reduction. Advances in Neural Information Processing Systems 15. 2003: 1959–1966.

15. He X, Yan S, Hu Y, Niyogi P, Zhang H. Face Recognition Using Laplacianfaces. IEEE Trans Pattern Anal Mach Intell. 2005; 27: 328–340. https://doi.org/10.1109/TPAMI.2005.55 PMID: 15747789

16. Han Y, Xu Z, Ma Z, Huang Z. Image classification with manifold learning for out-of-sample data. Signal Processing, 2013; 93: 2169–2177.

17. Choi SJ, Woods J.W. Motion-compensated 3-D subband coding of video. IEEE Trans Image Process. 1999; 8: 155–167. https://doi.org/10.1109/83.743851 PMID: 18267464

18. Deng C, He X, Han J, Zhang H. Orthogonal Laplacian faces for face recognition. IEEE Trans Image Process. 2006; 15: 3608–3614. PMID: 17076419
19. Li R, Zeng B, Liou M L. A new three-step search algorithm for block motion estimation. IEEE Trans Circuits Syst Video Technol. 1994; 4: 438–442.

20. Chang H, Yang H, Gan Y, Wang M. Sparse feature fidelity for perceptual image quality assessment. IEEE Trans Image Process. 2013; 22: 4007–4018. https://doi.org/10.1109/TIP.2013.2266579 PMID: 23751962

21. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annu Rev Neurosci. 2001; 24: 1193–216. https://doi.org/10.1146/annurev.neuro.24.1.1193 PMID: 11520932

22. Zhang L, Zhang L, Mou X, David Z. FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Trans Image Process. 2011; 20: 2378–2386. https://doi.org/10.1109/TIP.2011.2109730 PMID: 21292594

23. Masry M, Hemami S S, Sermadevi Y. A scalable wavelet-based video distortion metric and applications. IEEE Trans Circuits Syst Video Technol. 2006; 16: 260–273.

24. Sheikh HR, Wang Z, Cormack LK, Bovik AC. LIVE Image Quality Database, Sep. 2005, http://live.ece.utexas.edu/research/quality/subjective.htm.

25. Pinson MH, Wolf S. A new standardized method for objectively measuring video quality. IEEE Trans Broadcasting. 2004; 50: 312–322.

26. Amirshahi SA, Larabi M. Spatial-temporal Video Quality Metric based on an estimation of QoE. 2011 Third International Workshop on Quality of Multimedia Experience (QoMEX), 2011: 84–89.