Quality Assessment for Video with Degradation Along Salient Trajectories

Jinjian Wu, Yongxu Liu, Weisheng Dong, Guangming Shi, and Weisi Lin, Fellow, IEEE

***Abstract*—With the rapid growth of digital video through the Internet, a reliable objective video quality assessment (VQA) algorithm is greatly demanded for video management. Motion information plays a dominant role for video perception, and the human visual system (HVS) is able to track moving objects effectively with eye movement. Moreover, the middle temporal area of the brain is selective for moving objects with particular velocities. In other words, visual contents that along the motion trajectories will automatically attract our attention for dedicated processing. Inspired by the motion related process in the HVS, we suggest to analyze the degradation along attended motion trajectories for VQA. The characteristic of motion velocity along each trajectory is analyzed for temporal quality measurement. Meanwhile, visual information along each trajectory is extracted for joint spatial-temporal quality measurement. Finally, considering the spatial quality degradation from each frame, a novel Full-reference Assessor along Salient Trajectories (FAST) for VQA (which combines the spatial, temporal, and joint spatial-temporal quality degradations) is introduced. Experimental results on five publicly available VQA databases demonstrate that the proposed FAST VQA model performs consistently with the subjective perception. The source code of the proposed method will be available at http://web.xidian.edu.cn/wjj/paper.html.**

***Index Terms*—Video Quality Assessment, Motion Trajectory, Optical Flow, Spatial-Temporal Quality Degradation**

I. Introduction

With the rapid development of digital devices and Internet, videos have tremendously increased (which takes up more than 70% Internet traffic). In order to efficiently process such huge data, a powerful video management is urgently demanded. Meanwhile, digital videos suffer from distortions during several processing states (video compression, transmission, storage, and so on), with which the video quality will be severely degraded. Therefore, a faithful quality evaluator is greatly required for video management.

During the past decades, a large amount of video quality assessment (VQA) models have been introduced. According to the amount of the reference information, VQA methods can be classified into three categories: full-reference (FR) with the whole reference information available, reduced-reference (RR) with only limited number of features about the reference, and no-reference (NR) without any reference information when predicting the quality of the test video. Since FR-VQA is still a challenging task due to the obscurity of human subjective perception, we focus on FR-VQA in this work.

In the early stage, classical image quality assessment (IQA) models [1, 2] were directly adopted, and the video quality was measured with a frame-by-frame IQA procedure. As a natural way to calculate the signal error, the mean square error (MSE)/peak signal-to-noise ratio (PSNR) was often adopted for VQA [3]. Though such types of VQA models perform efficiently, they are correlated poorly with the subjective perception. Hence, some human visual system (HVS) based IQA models were adopted to measure the quality of each frame. Moreover, considering the effect of motion, the temporal information was simply incorporated as a weighting factor for VQA modeling, e.g., weighted SSIM and weighted VIF for VQA [4]. However, these models did not directly use motion information for temporal distortion measurement, which limited their accuracy for quality prediction.

Motion information plays an extremely important role during video perception. Though inundated with visual contents from the video, the HVS only focuses its attention on a fraction of the input contents for dedicated processing [5, 6]. Generally, moving objects will automatically attract our attention [7, 8]. In other words, the HVS is more sensitive to distortions on moving objects. Therefore, distortions on moving objects should be highlighted for temporal quality prediction [9]. In [10], the temporal variations in several continue frames were calculated, and then pooled with visual attention for motion distortion measurement. In [11], a long-range motion distortion was measured along horizontal and vertical direction respectively for VQA. In [12], considering the motion driven foveation during video perception, a novel contrast sensitive function was deduced, and then applied to measure the wavelet-based distortion visibility during VQA. In [13], the optical flow was calculated to represent motion for temporal quality assessment. In the recent, a new space-time texture was employed to capture the temporal distortion in [14]. Moreover, in [15], the motion information was directly extracted in the 3D local cube by compact representation of energy. Though these models have greatly improved the performance of VQA, there is still a large gap between the existing VQAs and the subjective perception.

It is still a great challenge to effectively describe the complex motion for VQA modeling. Motion information is an essential characteristic of video. Researches on cognitive neuroscience indicate that the brain contains specific areas for motion processing [16]. When perceiving a video sequence,

Jinjian Wu and Weisheng Dong are with State Key Laboratory of Integrated Services Networks, School of Artificial Intelligence, Xidian University, Xian, China. E-mail: jinjian.wu@mail.xidian.edu.cn.

Yongxu Liu and Guangming Shi are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, China.

Weisi Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore.

This work was partially supported by the NSF of China (Nos. 61772388, 61632019, 61621005, 61472301), and the Young Star Science and Technology Project (No. 2018KJXX-030) in Shaanxi province.

the visual stimuli are firstly processed in the primary visual cortex (V1), in which the amount of motion energy is efficiently measured [17]. Next, the middle temporal (V5) area receives the stimuli about motion from V1 for dedicated processing. For example, V5 will guide the eye movement on video sequences, which helps us to focus our attention on the moving objects for detailed feature extraction [16]. Moreover, some neurons in V5 are selective for moving objects with a particular velocity (both speed and direction) [18]. As a result, the HVS is highly adaptive to track moving objects with pursuit eye movements.

Inspired by the motion related process in the HVS (especially in V5), we suggest to directly extract the attended moving objects (both motion contents and their velocities) for motion quality prediction. As the paths of moving objects in a video, trajectories can effectively capture the motion information [19]. Thus, in this work, degradations along the motion trajectories are measured for VQA. As a convenient representation of motion, optical flow is firstly computed for motion trajectory searching. With the guidance of trajectories, the motion velocities (both speed and direction) are analyzed for temporal quality measurement. Meanwhile, visual features of motion content along the trajectories are extracted for spatial-temporal quality measurement. Finally, considering the spatial quality degradation, a novel Full-reference Assessor along Salient Trajectories (FAST) for VQA modeling is built.

The main contributions of this work are as follows:

1) Inspired by the attention mechanism on moving objects in HVS, a novel motion trajectory based dynamic degradation measurement is built. Comparing to other existing trajectory-based VQA methods, the proposed trajectory-based model concentrates only on moving objects for dynamic degradation, and covers a longer range, which conforms to the perception procedure of HVS to some degree.
2) Motivated by the motion-related perceptual processing mechanism in the human brain, the changes on both velocities and contents along motion trajectories are suggested to be thoroughly analyzed, which is also distinct from other existing VQA methods. As a result, motion degradation along trajectories are measured.
3) By incorporating the spatial, temporal, and joint spatial-temporal degradations, a novel FR-VQA method is introduced. The proposed FAST presents high correlation with the subjective perception. Additionally, experimental results also demonstrate the computational efficiency.

The rest of this paper is organized as follows. In Section II, trajectory extraction method for VQA is introduced. In Section III, the proposed trajectory-based VQA model is established. Experimental demonstrations are conducted In Section IV. Finally, Section V gives a conclusion of this work.

II. Trajectory Extraction

Motion plays a critical role in video viewing, and the paths of moving objects in successive video frames form motion trajectories. In a video sequence, each trajectory tracks a corresponding moving point over frames (along time). Therefore, trajectories represent the local motion information in a video, and the HVS is extremely sensitive to visual contents along trajectories. In this section, motion trajectories are estimated through points sampling and position prediction. To this end, the keypoints in the initial frame are firstly extracted. Next, the optical flow is computed with adjacent frames. And then, the position of each keypoint in the next frame is predicted. After predicting the position in each frame, the trajectory is expressed as a set of positions along time-axis.

A. Keypoints Selection

As a basic problem, keypoints selection is critical for subsequent procedure. A good point should be located in the major attended region, and can be tracked stably. To this end, a classic and widely-used method was proposed in [20], for which only points with large autocorrelation matrix eigenvalues are reserved. Since [20] is only dedicated on the stability of points in the tracking process, which makes the probability of points on the boundary of the frame equal to the one in the center. However, the HVS is more likely to focus attention on the center of the screen (comparing to the boundary). Thus, center bias mechanism should be incorporated to increase the probability of points located at the center of the scene, and also decrease the ones on the boundary.

Specifically, points on the initial frame of the distorted videos are densely sampled with a step size of 5 pixels, with the criterion suggested in [20]. Given a certain point, a 2D structure tensor (a 2×2 symmetric matrix) $\mathcal{Z}$ is computed firstly. And two eigenvalues λ_1 and λ_2 of $\mathcal{Z}$ are then calculated. The point is picked only when

$$\min(\lambda_1, \lambda_2) > T, \tag{1}$$

where T is a threshold, which is adaptive to the maximum value E_m among all the minimum eigenvalues of points in the frame. Here, T is set to $0.05\times E_m$.

Before Eq. (1), a weighting factor w for eigenvalues is firstly computed using a simple but effective center bias mechanism. Concretely, an distance map $\hat{w}$ is calculated as

$$\hat{w}(x,y) = \frac{(x-\frac{W_f}{2})^2 + (y-\frac{H_f}{2})^2}{(\frac{W_f}{2})^2 + (\frac{H_f}{2})^2}, \tag{2}$$

and normalized to $[0,1]$. Here, W_f and H_f represent the width and height of the frame, respectively. And w can be simply calculated as

$$w = 1 - \hat{w}. \tag{3}$$

A higher value in the factor map w means the location is more nearby the center of the frame, and more likely to be attended by our eyes (on account of center bias mechanism). Eigenvalues in Eq. (1) are tuned by w to help keypoints selection as interpreted above.

B. Optical Flow Computation

Optical flow is the motion distribution of two adjacent frames, which contains the velocity information of moving objects. Thus, optical flow is usually used to measure the

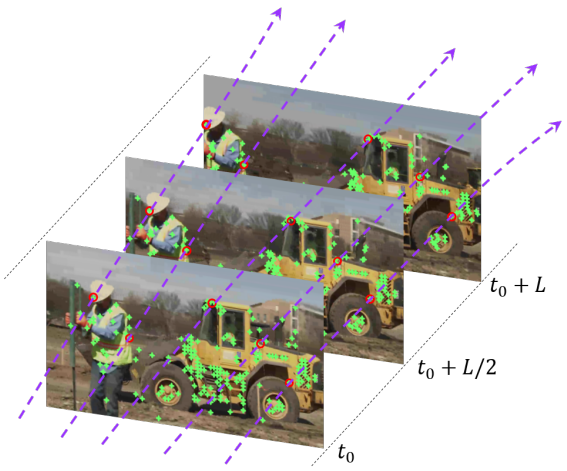


Fig. 1: Trajectory extraction in a video sequence

displacement of points during trajectory extraction. In this work, a dense optical flow algorithm [21] is adopted, which computes the displacement through a polynomial expansion to approximate each pixel neighborhood.

For a given point (x, y) in the first frame, its neighborhood, expressed as $\mathbf{x}$ in a local coordinate system, can be approximated by a polynomial as

$$f_1(\mathbf{x}) = \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + \mathbf{b}_1^T \mathbf{x} + c_1, \tag{4}$$

where $\mathbf{A}_1$ is a symmetric matrix, $\mathbf{b}_1$ is a vector, and c is a scalar. Considering that $\mathbf{d}$ is the offset for the neighborhood from the first frame to the second frame, $\mathbf{x}$ in the second frame can be approximated as

$$\begin{aligned} f_2(\mathbf{x}) &= f_1(\mathbf{x} - \mathbf{d}) \\ &= \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + (\mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d})^T \mathbf{x} + \mathbf{d}^T \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^T \mathbf{d} + c_1 \\ &= \mathbf{x}^T \mathbf{A}_2 \mathbf{x} + \mathbf{b}_2^T \mathbf{x} + c_2. \end{aligned} \tag{5}$$

By equating the coefficients, the offset can obtained as

$$\mathbf{d} = \begin{bmatrix} u \\ v \end{bmatrix} = -\frac{1}{2} \mathbf{A}_1^{-1} (\mathbf{b}_2 - \mathbf{b}_1), \tag{6}$$

where u and v are the horizontal and vertical components of displacement at point (x, y). Through solving a weighted least square problem in the neighborhood iteratively with multiple scales, the displacement at point (x, y) can be estimated, and more details are described in [21].

C. Trajectory Generation

Once a point is sampled from the first initial frame (i.e., $t = t_0$), denoted as $p^{t_0} = (x^{t_0}, y^{t_0})$ through Eq. (1), its corresponding position in the second frame (i.e., $t = t_1$) can be estimated as

$$p^{t_1} = (x^{t_1}, y^{t_1}) = (x^{t_0} + u, y^{t_0} + v), \tag{7}$$

where u and v are the horizontal and vertical components at position (x^{t_0}, y^{t_0}) of the optical flow computed using Eq. (6) from the first frame to the second frame, and (x^{t_1}, y^{t_1}) denotes the estimated position in the second frame.

By repeating the procedure above from t_0-th frame to $t_{\mathcal{L}}$-th frame, the point is tracked in a consecutive time, and a trajectory is created based on the combination of these points, which can be denoted as

$$\mathcal{T} = \{p^{t_0}, p^{t_1}, ..., p^{t_{\mathcal{L}}}\}. \tag{8}$$

An example of trajectories is shown in Fig. 1. These sampled points are marked with green cross at t_0. In the subsequent frames, these points are estimated with position prediction using optical flow. In order to give a vivid example for trajectory extraction, some points are labeled as red circles, and their corresponding trajectories are labeled as purple dash.

Trajectories generated from Eq. (7) are rough, and generally not all of them should be considered for further modeling on account of subjective and objective reasons. As for objective aspects, displacement from optical flow algorithms can be erroneous and the trajectories extracted may be invalid. Thus, trajectories with random steps or out of frame boundary are removed, so as those containing unusual large displacement between two adjacent frames. For the subjective aspect, trajectories are used for dynamic degradation modeling. So these trajectories which are totally static should be excluded. Meanwhile, if the Euclidean distance of two trajectories are small enough, they may share much common information along trajectories. Such kind of trajectories are recommended to reserve only a representative one (removing the another).

In our implementation, the standard deviation of point positions along a trajectory is computed, and if the standard deviation is larger than $maxStd = 10 \times \frac{min(W_f, H_f)}{256}$ or any displacement between two adjacent frames is beyond $maxDis = 10 \times \frac{min(W_f, H_f)}{256}$, the trajectory is discarded as an abnormal one. The thresholds are adaptive to the size of the frame due to varied resolutions in reality. Similarly, the sum of displacements along the trajectory is calculated, and it is believed to be a static one if the sum is less than $minDis = 1$. Meanwhile, the threshold for the decision on two repeated trajectories is to guarantee the diversity of each trajectory, and set to $threRep = 0.8 \times \mathcal{L} \times \frac{W}{2}$, where W is related to the step of feature extraction, which will be illustrated in III-A.

Besides, the length of trajectories is important to some extent. For a short trajectory, it contains limited motion information; while a long one has a higher risk of drifting from its original position. In this work, the length $\mathcal{L}$ is set as 18 to obtain a good trade-off (detailed experimental analysis and will be given in Section IV-E). Noted that from t_0-th to $t_{\mathcal{L}}$-th frame, not a single trajectory is generated since many points are sampled at t_0-th frame. Moreover the trajectory generation procedure indicates latently that the video is divided into many subsequences (each subsequence contains $\mathcal{L}+1$ frames). And considering the possible removal of trajectories and variation of video content, points are sampled every $\frac{\mathcal{L}}{2}$ frames without overlapping previous points. This step ensures a reasonable overlapping for video sequences in time domain, but non-overlapping for video content in spatial domain.

Further, in this paper, points are sampled on the distorted video, and trajectories are also generated using the distorted optical flow. It is believed to be more reasonable since different distorted videos corresponding to the same reference may

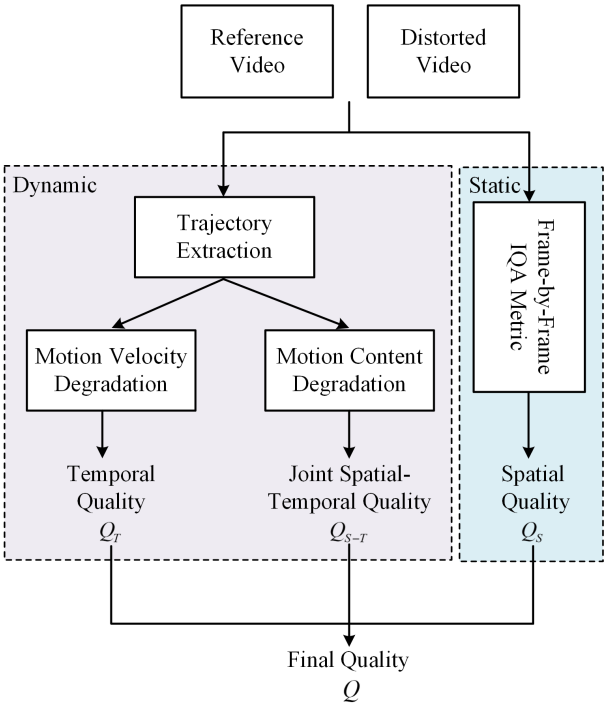


Fig. 2: Overview of the proposed VQA model

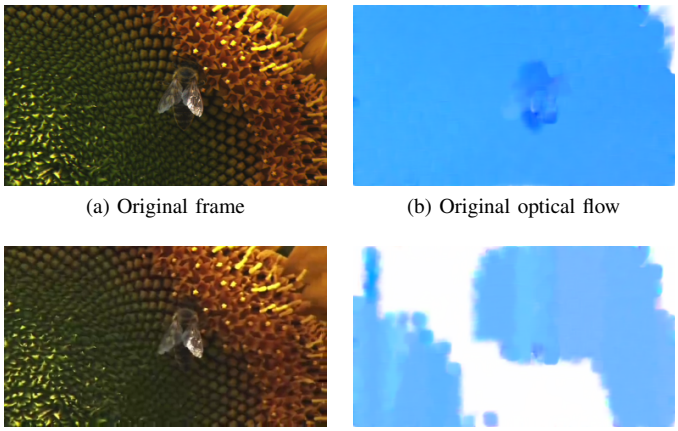
result in different attended motion informations. And in real life, subjects are directly faced with distorted videos for quality perception, rather than the reference video (a further experimental analysis is given in Section IV-E).

III. VQA Modeling

In this section, a novel trajectory degradation based VQA model is built. Since distortions on the moving objects will generate obvious quality degradation, motion information along trajectories, including motion velocity and motion content, is extracted to measure the degradation. Firstly, motion velocity is mapped into histograms to measure the temporal quality. Then, motion content is extracted to calculate the joint spatial-temporal quality. Moreover, considering the static degradation on each frame, the spatial quality is also estimated. Finally, Combing the spatial, temporal, and spatial-temporal components, the quality of a video is predicted. An overview of the proposed model is shown in Fig. 2.

A. Temporal Quality Measurement

Velocity is one of the most important factor of motion, which includes both speed and direction, and temporal distortions in videos always change the velocity. Thus, degradations on motion speed and direction are calculated for temporal quality measurement. In this paper, the motion speed and direction are obtained with the optical flow algorithm mentioned above. Optical flow can depict the motion information well [9, 13], as well as degradations on motion velocity. An intuitive illustration is given in Fig. 3. The original frame is shown in Fig. 3 (a), and Fig. 3 (b) is the corresponding optical flow map between the previous frame to the current one. Fig. 3 (c) is distorted by H.264 compression, which contains



(a) Original frame (b) Original optical flow

(c) Distorted frame (d) Distorted optical flow

Fig. 3: An illustration for the degradation on optical flow

texture floating due to camera motion. If we only compare the two individual frames, there are only tiny difference (a little blur in Fig. 3 (c)). However, the optical flow can effectively capture the texture floating, as shown in Fig. 3 (d), which is extremely annoying during perception. Since the optical flow can effectively capture distortions on motion velocity, the temporal quality is suggested to be measured by degradations on motion speed and direction along trajectories. A flow chart of the temporal degradation estimation is shown in Fig. 4.

Specifically, for each point p_j in trajectory $\mathcal{T}_i$, the local region centered at p_j with a size of $W \times W$ is extracted on its corresponding optical flow map. Then, an optical flow tube $\mathcal{C}_{\mathcal{T}_i}$ along the trajectory is acquired, whose size is $W \times W \times \mathcal{L}$. Though there are $\mathcal{L}+1$ points, we can only use the preceding $\mathcal{L}$ points. That is because $\mathcal{L}+1$ frames video sequence can only produce $\mathcal{L}$ frames optical flow. The tube $\mathcal{C}_{\mathcal{T}_i}$ contains motion velocity (magnitude and orientation) of the local regions along $\mathcal{T}_i$, and any motion change in these regions between the reference and distorted video can be captured.

Let $d = (u, v)$ denotes the value of the optical flow at p_j, the motion speed $\mathcal{S}$ and direction θ can be calculated as

$$\begin{aligned} \mathcal{S} &= \sqrt{u^2 + v^2} \\ \theta &= \arctan\frac{v}{u}. \end{aligned} \tag{9}$$

Then, according to its spatial location, the tube is divided into four portions with a size of $\frac{W}{2} \times \frac{W}{2} \times \mathcal{L}$ (each is marked with different colors as shown in Fig. 4). This division can be considered as a finer representation, since the procedure would result in four locational cuboids (the upper left, upper right, lower left and lower right) along the trajectory, rather than the whole massive tube. By mapping the motion speed of each element in every cuboid to a $\mathcal{K}$ bins histogram based on the motion direction, a $2 \times 2 \times \mathcal{K}$ histogram is developed for the motion velocity representation.

Noted that this representation is based on the original histogram of optical flow (HOF), two aspects should be pointed out: 1) The whole cuboid of the optical flow with multiple frames along the trajectory is considered as a good representation of motion flow in a consecutive time. However,

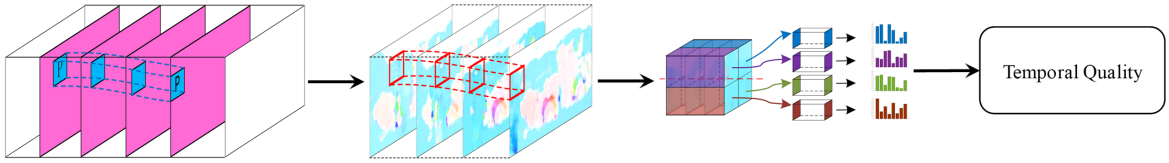


Fig. 4: Flow-chart of the temporal degradation measurement along trajectories

the original HOF is based on a single frame of optical flow, in which the motion is isolated for every single frame. 2) Spatial locational division makes the histogram more representative for the tube along the trajectory since more location-related details can be reserved.

Along the same trajectory $\mathcal{T}_i$, the histogram is computed for the reference and the distorted video as $\mathcal{H}_i^r$ and $\mathcal{H}_i^d$. The dissimilarity between $\mathcal{H}_i^r$ and $\mathcal{H}_i^d$ reflects the difference between the reference and the distorted video in motion velocity, which is calculated as

$$\mathcal{DV}_i = 1 - \frac{1}{M}\sum_{m=1}^{M}\frac{2\cdot\mathcal{H}_{i,m}^r\cdot\mathcal{H}_{i,m}^d + C_1}{(\mathcal{H}_{i,m}^r)^2 + (\mathcal{H}_{i,m}^d)^2 + C_1}, \tag{10}$$

where $M = 2\times 2\times\mathcal{K}$ is the length of the histogram, and C_1 is set to 0.00001 (a small constant to avoid instability). $\mathcal{DV}_i$ characterizes the dissimilarity of motion velocity along the i-th trajectory $\mathcal{T}_i$, and ranges from 0 (two histograms are wholly identical) to 1.

For each trajectory in the k-th video subsequence, the dissimilarity of motion velocity can be calculated. The whole temporal quality of the subsequence can be estimated using two simple statistical indicators, i.e., the mean value and the standard deviation. Even with the same mean value of frame-by-frame spatial quality, the final quality of a whole video can be quite different due to the diverse standard deviations [22]. Thus, besides the mean value, the standard deviation value can also benefit the quality prediction. In this work, the sum of both the mean value and deviation value is suggested for a simple and effective pooling method. Let $\mathcal{DV}$ denotes $\{\mathcal{DV}_1, \mathcal{DV}_2, ..., \mathcal{DV}_N\}$, where N is the number of trajectories in the video subsequence, the temporal quality is computed as

$$\begin{aligned}\mathcal{Q}_T^k &= \mathrm{Mean}(\mathcal{DV}) + \mathrm{Stdev}(\mathcal{DV})\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{DV}_i + \sqrt{\frac{\sum_{i=1}^{N}(\mathcal{DV}_i - \frac{1}{N}\sum_{i=1}^{N}\mathcal{DV}_i)^2}{N-1}}.\end{aligned} \tag{11}$$

B. Spatial-Temporal Quality Measurement

Besides the motion velocity, the visual information of moving objects (i.e., the motion content) also plays a dominant role in visual perception. Thus, the degradation on the motion content is measured for the joint spatial-temporal quality estimation. A flow chart of the joint spatial-temporal degradation estimation is shown in Fig. 5.

Firstly, the motion content is extracted along the motion trajectories. Similar to the procedure of extracting motion velocity above, a tube of motion content is extracted from the subsequence (rather than the optical flow). Since the HVS is highly sensitive to changes on the visual content, the motion content changes along x, y, and t directions are calculated for visual content representation.

In this work, the content changes in the tube is calculated with 3-D filters. Fig. 6 shows a 3-D filter along the x-axis, which is denoted as dx. Meanwhile, the filters for y-axis (dy) and t-axis (dt) can be obtained by simply rotating dx with 90° along t-axis and y-axis, respectively.

For a given tube with motion content V along the motion trajectory $\mathcal{T}_i$, its visual content is represented as

$$\mathcal{M}_{ST} = \sqrt{Comp_x^2 + Comp_y^2 + Comp_t^2}, \tag{12}$$

where the component $Comp_k = V * dk$, dk is the filter along k direction (e.g., x, y, and t), and $*$ is convolution operation.

The visual content is computed for the reference as well as the distorted video, noted as $\mathcal{M}_{ST}^r$ and $\mathcal{M}_{ST}^d$, respectively. And the dissimilarity between them can be expressed by the standard deviation of the similarity for each element. Specifically, we denote $\mathcal{DC}_i$ as the dissimilarity of the motion content along the trajectory $\mathcal{T}_i$ between the reference and the distorted, and

$$\mathcal{DC}_i = \mathrm{Stdev}(\mathcal{S}_{im}), \tag{13}$$

where $\mathcal{S}_{im}$ is of the same size of $\mathcal{M}_{ST}^r$ and $\mathcal{M}_{ST}^d$, and defined as

$$\mathcal{S}_{im}(x,y,t) = \frac{2\cdot\mathcal{M}_{ST}^r(x,y,t)\cdot\mathcal{M}_{ST}^d(x,y,t) + C_2}{(\mathcal{M}_{ST}^r(x,y,t))^2 + (\mathcal{M}_{ST}^d(x,y,t))^2 + C_2}. \tag{14}$$

As illustrated in [23], C_2 can work as a gradient masking parameter together with avoiding the instability of Eq. (14), which is set to 255 corresponding to the maximum of the 8-bit luminance.

Once the dissimilarity of motion content along all the trajectories in the k-th subsequence is calculated, the joint spatial-temporal quality in the subsequence, similar to Eq. (11), can defined as

$$\begin{aligned}\mathcal{Q}_{ST}^k &= \mathrm{Mean}(\mathcal{DC}) + \mathrm{Stdev}(\mathcal{DC})\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{DC}_i + \sqrt{\frac{\sum_{i=1}^{N}(\mathcal{DC}_i - \frac{1}{N}\sum_{i=1}^{N}\mathcal{DC}_i)^2}{N-1}},\end{aligned} \tag{15}$$

where $\mathcal{DC} = \{\mathcal{DC}_i, i = 1, 2, \ldots, N\}$, and N is the number of trajectories in the k-th video subsequence.

C. Spatial Quality Measurement

Similar to many other VQA works, the frame-by-frame quality degradation is also measured with existing IQA algorithms [24–27], and the mean value of all frames is seen as the spatial quality degradation. Since the temporal degradation and

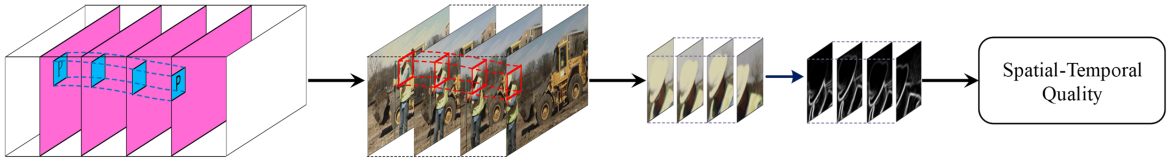


Fig. 5: Flow-chart of the spatial-temporal degradation measurement along trajectories

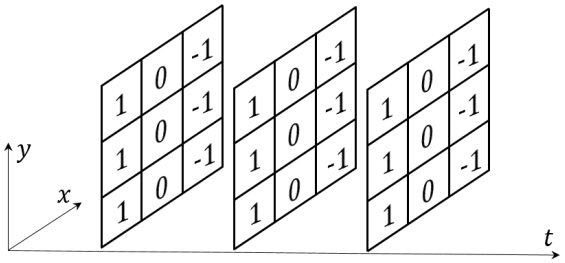


Fig. 6: 3-D operator for x direction – dx

the joint spatial-temporal degradation are computed in each video subsequence, the spatial quality is also generated in each subsequence. Considering a trade-off of computational cost and performance effectiveness, GMSD [28] is adopted here to compute the spatial quality of each frame, and the spatial quality in the k-th subsequence is computed as

$$\mathcal{Q}_S^k = \frac{1}{\mathcal{L}} \sum_{j=1}^{\mathcal{L}} GMSD_j \tag{16}$$

D. Combination

A good combination among these spatial, temporal, and joint spatial-temporal components would definitely improves the accuracy of quality estimation. However, it's hard to guess how to combine these into a total quality in human brain. As a widely-used method, the total quality in k-th subsequence can be simply expressed by the multiplication of the three components:

$$\mathcal{Q}^k = \mathcal{Q}_S^k \cdot \mathcal{Q}_T^k \cdot \mathcal{Q}_{ST}^k. \tag{17}$$

And we do believe that a good temporal pooling strategy would benefit a lot, but for simplicity, in this work the final quality of the test video is obtained by a mean value of all subsequences:

$$\mathcal{Q}_{\text{final}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{Q}^k, \tag{18}$$

where K is the total number of subsequences.

IV. Experimental Results and Analysis

In this section, a brief description of databases and protocol is firstly given. Then, the efficiency of each component (i.e., S: spatial degradation, T: temporal degradation, and S-T: spatial-temporal degradation) in our FAST VQA model is thoroughly analyzed. Next, a comprehensive comparison between the proposed FAST model and the existing VQA methods is demonstrated. Moreover, the comparison of computational complexity for VQA methods is conducted. Finally, some parameter settings in our method are analyzed experimentally.

A. Database and Protocol

In order to verify the performance of the proposed VQA model, five databases with different frame sizes ranging from VGA to HD are adopted, i.e., the IVC-IC video database [29] with a resolution of 640×480, the LIVE video database [30] with a resolution of 768×432, the CSIQ video database [11] with a resolution of 832×480, the MCL-V database [31] with a resolution of 1920×1080, and the IVP video database [32] with a resolution of 1920×1088.

A brief list of databases is shown as Tab. I. The IVC-IC video database contains 60 reference videos, and each is degraded with four randomly selected distortions from twenty different degradations, resulting 240 distorted videos. The LIVE video database contains 150 distorted videos and 10 reference videos, with four distortion types: MPEG-2 compression distortion, H.264 compression distortion, and transmission distortions over IP networks and wireless networks based on H.264 compression. The CSIQ video database contains 216 distorted videos corresponding to 12 reference videos with six distortion types, including: H.264 compression, HEVC/H.265 compression, Motion JPEG compression, Wavelet-based compressing using the Snow codec, simulated wireless transmission loss based on H.264 compression, and additive white noise. The IVP video database contains 10 reference videos and 128 distorted videos with four distortion types: H.264 compression, MPEG-2 compression, Dirac-wavelet compression, and transmission error over IP network based on H.264 compression. The MCL-V database contains 12 reference videos with two distortion types: H.264/AVC compression and compression followed by scaling (or simply called scaling), and four distortion levels, thus creating 96 distorted videos.

Three widely-used metrics are adopted in this work for performance comparison, which are the Pearson linear correlation coefficient (PLCC), the Spearman rank order correlation coefficient (SRCC), and the root mean squared error (RMSE). PLCC and SRCC measure the correlation between the predicted quality and the ground truth, which ranges from 0 to 1; and RMSE denotes the relative error. A better VQA method will result in higher PLCC and SRCC values, and a lower RMSE value.

B. Efficiency Analysis

To give an intuitive efficiency analysis for each component (i.e., S, T, and S-T) in our model, four videos (named rh_10, rh_5, rh_6, and rh_12) on LIVE video database are shown in Fig. 7, where rh_5 is distorted by wireless distortion, rh_6 is distorted by IP network distortion, and the other two are distorted by H.264 compression distortion. Since

TABLE I: Information About the Five VQA Databases

Database	Resulotion	Ref. No.	Dist. No.	Distortion Type
IVC-IC	640 × 480	60	240	Four random distortions
LIVE	768 × 432	10	150	MPEG-2, H.264, IP, Wireless
CSIQ	832 × 480	12	216	H.264, Packet Loss, MJPEG, Wavelet, White Noise, HEVC
IVP	1920 × 1088	10	128	H.264, MPEG-2, Dirac-wavelet, IP
MCL-V	1920 × 1080	12	96	H.264, Scaling

(a) rh_10 (b) rh_5

(c) rh_6 (d) rh_12

Fig. 7: An intuitive example on efficiency analysis

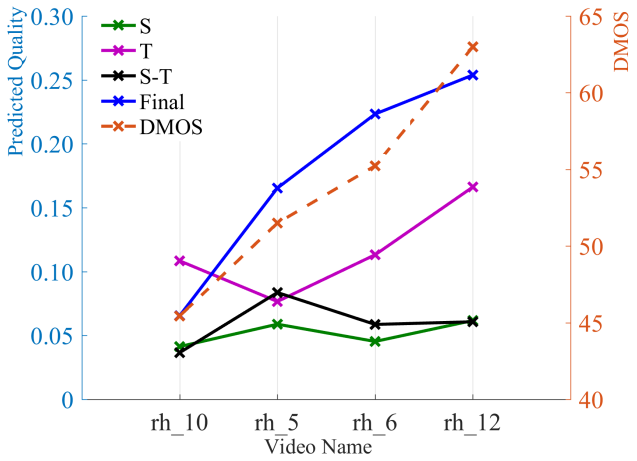


Fig. 8: Predicted quality and DMOS for Fig. 7

rh_5 and rh_6 are distorted through network, the artifacts are easier to be observed (red rectangles in Fig. 7), while rh_12 always represents transient temporal distortion due to motion compensation mismatch, which is harder to be recognized in static status. The DMOS for $(a) \sim (d)$ are 45.4363, 51.4980, 55.2291, and 62.9934, respectively. Moreover, a lower DMOS value means a higher perceptual quality. Thus, rh_10 possesses the best perceptual quality, and rh_12 looks worst.

We test our algorithm on the four videos, and the qualities of components are shown in Fig. 8 (as well as the final combing quality). Meanwhile, the final quality values are rescaled for a better visualization by multiplying a proper constant, which would not disturb the tendency of the curve. As can be seen in Fig. 8, single component can hardly keep the consistency with human perception, but they are complementary to some extent. For example, for rh_5, the component T (the purple line) performs poorly (returns a high quality value, which conflicts with the ground truth), and the components S and S-T perform accurately. As a result, the components S and S-T can amend the limitation of T on rh_5, and their combination result returns a right quality score. Such situation also happens for the other three video sequence, and finally accurate quality scores are obtained for them (shown as the blue line, which is highly consistent with DMOS, the brown dash line). Therefore, the three components (i.e., S, T, and S-T) are very complementary, and their combination can return a better quality prediction result than any single component.

TABLE II: Efficiency Analysis of Each Component on LIVE Video Database

Crit	S	T	S-T	Final
PLCC	0.7441	0.6350	0.7975	0.8892
SRCC	0.7297	0.6158	0.7853	0.8800
RMSE	7.3337	8.4801	6.6232	5.0221

Besides the special case given above, a statistical analysis of these components on LIVE video database is conducted. The performances of these components are listed on Tab. II. As can be seen from Tab. II, arbitrary single component can not perform well on the database (with lower PLCC and SRCC values, and higher RMSE values), but their combination can greatly improve the prediction accuracy (returns the highest PLCC and SRCC, and the lowest RMSE). Therefore, the degradation from the spatial (S component), temporal (T component), and spatial-temporal (S-T component) should be jointly considered for VQA.

C. Performance Comparison

To comprehensively illustrate the efficiency of the proposed FAST method, nine state-of-art FR VQA models (i.e., MOVIE [9], VQM_VFD [33], stRRED [34], Vis3 [11], FePVQ [35], FLOSIM [13], VMAF [36], SpEED [37] and Peng [14]), as well as two FR IQA models (GMSD [28] and MS-SSIM [38]) are compared with our model on five VQA databases (i.e., LIVE, CSIQ, IVP, MCL-V, and IVC-IC). As for MOVIE, to reduce the computation cost, videos on IVP and MCL-V are usually downsampled from 1920 × 1088 to 960×544, or from 1920×1080 to 960×540 with ffmpeg using

TABLE III: VQA Performance Comparison on Five VQA Databases

Database	Crit.	VQA										IQA	
		Proposed	Peng	SpEED	VMAF	FLOSIM_fb	FePVQ	Vis3	stRRED	VQM_VFD	MOVIE	GMSD	MS-SSIM
LIVE (150)	PLCC	**0.8892**	0.8235	0.7816	0.7615	**0.8421**	0.8326	0.8336	0.8111	0.7853	0.8113	0.7371	0.7431
	SRCC	**0.8800**	0.8216	0.7744	0.7545	**0.8389**	0.8279	0.8168	0.8007	0.7736	0.7884	0.7262	0.7364
	RMSE	5.0221	6.2269	6.8468	7.1157	5.9200	-	6.0635	6.4221	6.7966	6.4183	7.4184	7.3463
CSIQ (216)	PLCC	**0.8850**	0.7741	0.7368	0.6243	0.7264	0.8210	0.8223	0.7933	**0.8388**	0.7924	0.8214	0.7571
	SRCC	**0.9076**	0.7868	0.7423	0.6151	0.7318	0.8100	0.8326	0.8129	**0.8480**	0.8083	0.8409	0.7565
	RMSE	7.7417	10.526	11.243	12.990	11.428	-	9.4617	10.123	9.0530	10.144	9.4844	10.863
IVP (128)	PLCC	**0.8953**	0.6593	0.7820	0.5919	0.5443	**0.9110**	0.7959	0.7287	0.8466	0.8816	0.6838	0.5953
	SRCC	**0.9090**	0.6573	0.7934	0.5799	0.5450	0.8840	0.7948	0.7374	0.8494	**0.8844**	0.6860	0.5799
	RMSE	0.4709	0.7948	0.6592	0.8493	0.8204	-	0.6400	0.7237	0.5626	0.4991	0.7714	0.8495
MCL-V (96)	PLCC	**0.7816**	0.7469	0.7440	0.7792	0.5734	-	0.6470	0.7548	**0.8096**	0.7763	0.6501	0.6462
	SRCC	**0.7863**	0.7319	0.7851	0.7766	0.5919	-	0.6353	0.7433	**0.8033**	0.7772	0.6449	0.6306
	RMSE	1.3845	1.4757	1.4893	1.3910	1.8181	-	1.6921	1.4558	1.3027	1.4002	1.6862	1.6936
IVC-IC (240)	PLCC	**0.9346**	0.9315	0.8788	**0.9435**	0.8810	-	0.9265	0.8134	0.9335	0.9076	0.9264	0.9122
	SRCC	**0.9365**	0.9258	0.9002	**0.9333**	0.8762	-	0.9177	0.8972	0.9226	0.9005	0.9196	0.9065
	RMSE	0.3886	0.3953	0.5290	0.3604	0.5143	-	0.4091	0.6365	0.3897	0.4565	0.4093	0.4455
Average	PLCC	**0.8771**	0.7871	0.7846	0.7401	0.7134	-	0.8051	0.7803	**0.8428**	0.8338	0.7638	0.7308
	SRCC	**0.8839**	0.7847	0.7991	0.7319	0.7168	-	0.7994	0.7983	**0.8394**	0.8318	0.7635	0.7220
Weighted Average	PLCC	**0.8897**	0.8077	0.7938	0.7543	0.7462	-	0.8301	0.7879	**0.8543**	0.8410	0.7955	0.7616
	SRCC	**0.8972**	0.8070	0.8066	0.7456	0.7479	-	0.8257	0.8154	**0.8512**	0.8395	0.7964	0.7544

its default setting. Also, considering the huge computational cost of Black and Anandan (BA) optical flow algorithm, Farneback optical flow algorithm is adopted for FLOSIM (denoted as FLOSIM_fb) and it has been demonstrated in [13] that using Farneback optical flow algorithm can still achieve a comparable performance. The parameters are followed as the author suggested, and we will give an intuitive comparison about computational cost in Sec. IV-D to demonstrate this substitution is reasonable and necessary. Meanwhile, the latest 0.6.1 version of VMAF is used here.

The performance comparison is listed in Tab. III, and the best two performances of PLCC and SRCC are highlighted. As is shown in Tab. III, comparing to these existing VQA methods, the proposed trajectory-based VQA model can reach a remarkable improvement. The performances (SRCC) on CSIQ, IVP and IVC-IC are pretty well (over 0.9), which demonstrates the high consistency with the subjective perception. Meanwhile, the results on LIVE and CSIQ obtain an improvement of 0.04 in terms of PLCC and SRCC beyond the second best algorithm, showing a large gap among other FR VQA algorithms. Moreover, from Tab. III, the proposed FAST shows a more stable performance for different video databases (ranging from VGA to HD), while other methods may exhibit some certain bias. For example, FePVQ performs well on IVP, but poorly on the other databases. A similar situation can be seen on MCL-V and IVC-IC video databases. Although VQM_VFD achieves the best performance on MCL-V (so does VMAF on IVC-IC), the proposed FAST method still possesses a competitive performance (ranking the second place and outperforming the others a lot). From this point of view, it is obvious that our proposed model exhibits a more reliable performance and presents a better performance of generalization. Meanwhile, the average and weighted average performances (listed at the bottom of Tab. III) also confirm the efficiency and stability of our proposed model (possessing the highest PLCC and SRCC values, and giving a great advantage over other VQA methods).

Furthermore, to give a detailed illustration for the efficiency, the performance (SRCC) on each individual distortion type of four video database is given in Tab. IV. Noted that distortion types on IVC-IC database are not considered in this procedure since they are randomly selected from twenty distortions. As it can be seen in Tab. IV, on each distortion type, the proposed VQA model still possesses an apparent advantage over other methods. As for several distortion types, like packet loss through network, as well as MPEG-2 compression distortion, the proposed FAST method is preferable to predict the perceptual effects of distortions. Moreover, FAST still performs a relatively favorable results among the other distortions. For example, as for H.264 compression distortion, there are four separate VQA models ranking the best on the four databases (i.e., FLOSIM_fb on LIVE, Peng on CSIQ, SpEED on IVP, and VQM_VFD on MCL-V). However, our method can always achieve competitive performances, and outperform the other methods on this distortion type, which, similarly, proves the good stability and generalization of the proposed model.

The average performance (SRCC) of each VQA models on 16 distortion types among the four databases is given at the bottom of Tab. IV, as well as the hit-count of the best performance on each individual distortion. Both of the average performance and the hit-count on individual distortion, give strong evidences to illustrate the preferable efficiency and great superiority of our method.

Additionally, the robustness of the proposed FAST method with Different IQA metrics for the spatial component is thoroughly investigated. In this paper, eight existing IQA metrics (i.e., MS-SSIM [38], VIF [39], IW-SSIM [40], FSIM [32], VSI [41], GMSD [28], MDSI [42], and PSIM [43]), spanning from the year of 2003 to 2017, are adopted for the frame-

TABLE IV: Performance (SRCC) Comparison on Each Individual Distortion Type

Distortion Type		Proposed	Peng	SpEED	VMAF	FLOSIM_fb	FePVQ	Vis3	stRRED	VQM_VFD	MOVIE
LIVE	Wireless	**0.9026**	0.7767	0.8045	0.7996	0.8113	0.8073	0.8394	0.7857	0.6919	0.8109
	IP	**0.8042**	0.6859	0.7664	0.6903	0.7798	0.7417	0.7918	0.7722	0.7271	0.7157
	H.264	0.8814	0.8570	0.7895	0.7462	**0.8917**	0.8725	0.7685	0.8195	0.7304	0.7644
	MPEG-2	**0.8315**	0.7521	0.6553	0.7099	0.8016	0.7513	0.7362	0.7193	0.8223	0.7613
CSIQ	H.264	0.9560	**0.9828**	0.9640	0.9284	0.9382	0.9120	0.9194	0.9768	0.9344	0.8960
	H.264 PL	**0.9359**	0.8468	0.8525	0.7701	0.9060	0.8730	0.8533	0.8546	0.8232	0.8677
	MJPEG	**0.9382**	0.8932	0.0713	0.8871	0.7591	0.8870	0.7349	0.7290	0.8896	0.8855
	Wavelet	0.9362	0.9318	0.9403	0.8965	0.7333	0.8750	0.8999	**0.9459**	0.8644	0.9012
	White Noise	0.9287	0.9230	0.9091	0.8831	0.8481	0.8480	0.9179	**0.9305**	0.8888	0.8245
	HEVC	0.9485	**0.9555**	0.8754	0.9287	0.9017	0.9410	0.9174	0.9099	0.9516	0.9349
IVP	Dirac	0.9159	0.7976	0.8594	0.9017	0.8309	0.8670	0.9155	0.8527	**0.9172**	0.8879
	H.264	0.8822	0.8246	**0.8841**	0.8653	0.6129	0.8560	0.8426	0.8655	0.8672	0.8482
	MPEG-2	**0.9462**	0.6685	0.7059	0.7922	0.7646	0.8560	0.7940	0.6912	0.8647	0.8162
	IP	0.8046	0.4745	0.7750	0.3103	0.5643	0.8190	0.7438	0.6672	0.7510	**0.8582**
MCL-V	H.264	0.7858	0.7415	0.8054	0.7939	0.6247	-	0.5868	0.7716	**0.8108**	0.7596
	Scaling	0.7747	0.7123	0.7418	0.7524	0.5548	-	0.6857	0.7040	**0.7955**	0.7783
Average		**0.8858**	0.8015	0.7750	0.7910	0.7702	-	0.8092	0.8122	0.8331	0.8319
Hit Count		6	2	1	0	1	0	0	2	3	1

TABLE V: Performance of VQA Modeling from Different Spatial Metrics

IQAs	LIVE		CSIQ		IVP	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
MS-SSIM	0.8914	0.8808	0.8814	0.8943	0.8410	0.8912
VIF	0.8633	0.8540	0.8616	0.8651	0.8574	0.8893
IW-SSIM	0.8894	0.8789	0.8815	0.8951	0.8389	0.8923
FSIM	0.8889	0.8804	0.8815	0.8943	0.8874	0.8916
VSI	0.8917	0.8805	0.8801	0.8926	0.8869	0.8917
GMSD	0.8892	0.8800	**0.8850**	**0.9076**	**0.8953**	**0.9090**
MDSI	**0.8965**	**0.8830**	0.8807	0.9019	0.8879	0.8949
PSIM	0.8860	0.8799	0.8796	0.8915	0.8870	0.8915

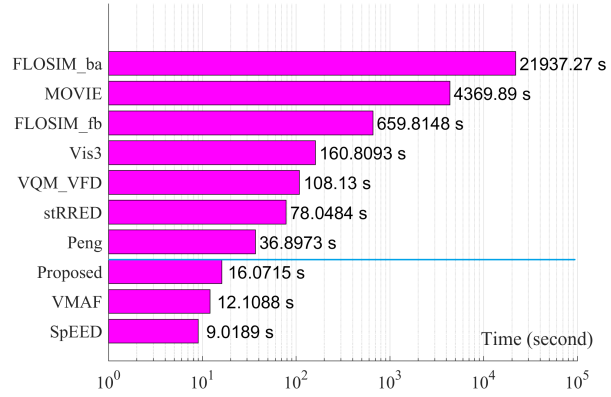


Fig. 9: Running time among VQA models

by-frame spatial quality assessment, and their performances on three VQA databases (i.e., LIVE, CSIQ, and IVP) are listed in Tab. V. As can be seen, the selection of IQA metrics for the spatial component is not critical for the proposed FAST method. Even though with different IQA metrics, the performances are quite similar. For example, the performance differences from FSIM, VSI, GMSD, MDSI and PSIM are no more than 0.01 (SRCC or PLCC value). This result suggests implicitly that the video quality is mainly related to motion degradation in most cases, rather than the statistic portion. Even though GMSD is not an ideal spatial quality estimator for optimization, considering the computational simplicity, GMSD is employed for the spatial quality in this work, which can be substituted flexibly with other preferable one based on the practical situation.

D. Effectiveness Comparison

Besides performance comparison, considering the limitation of computational cost in practical application, an experiment about running time is also conducted. Algorithms are running on an operation system of Windows 10 Enterprise (x64), with a 3.60 GHz Inter Core i7-4790 CPU and 16GB RAM. MOVIE is tested in a released model of Visual Studio 2013 Community, as well as VMAF, and the other VQA models are running with MATLAB R2016a. Our proposed method is implemented in C++ with OpenCV, and tested in MATLAB through hybrid programming.

Ten different distorted videos corresponding to the same reference video named $pa1\_25fps.yuv$ on the LIVE video database are selected. All test videos are with a resolution of 768×432 and possess 250 frames. And the average running time on these test videos is given as Fig. 9. As it can be seen, the proposed model is much effective than most of VQA methods. Existing models with a preciser predicted performance always need more computation cost (such as MOVIE, Vis3, and VQA_VFD), while a much faster algorithm (like SpEED, or VMAF) always could not perform well on all evaluating databases. As a balance of both, our trajectory-based method can make a more conspicuous performance, as

well as relatively less computation cost, which demonstrates the effectiveness and efficiency of our method, and shows a great advantage over other VQA models.

Furthermore, as it also can be seen in Fig. 9 that, FLOSIM implemented with BA optical flow algorithm (named FLOSIM_ba) needs more than six hours for a standard test video, while FLOSIM_fb onlys takes about eleven minutes. It is a hard work for FLOSIM_ba to accomplish the validation on the five databases on account of the huge computational cost, but much easier for FLOSIM_fb. From another point of view, since our VQA method is implemented with Farneback optical flow algorithm, using FLOSIM_fb rather than FLOSIM_ba seems more fair for comparison to some extent. Anyway, putting aside this view, and considering the major computational complexity, FLOSIM_fb is a better choice, reasonably and necessarily.

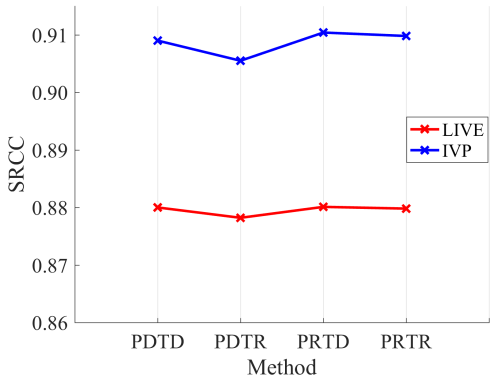


Fig. 10: Performance (SRCC) comparison with different method for points selection and trajectory generation on LIVE and IVP databases

E. Analysis on Parameter Settings

In this step, some options and parameter settings are experimentally analyzed, which contain options about points selection and trajectory generation, determination of the trajectory length $\mathcal{L}$, choice of the local region size W in motion velocity and motion content extraction, and selection of the bins $\mathcal{K}$ for motion velocity representation. On each experimental procedure, only single variable is controlled, while the others are fixed.

Firstly, in Section II, points are sampled from the distorted video frame, and trajectories are generated with the distorted optical flow by default. Actually, there are four types for this trajectory generation procedure: points sampled from the distorted frame and tracking points with the distorted optical flow (PDTD), points sampled from the distorted frame and tracking points with the reference optical flow (PDTR), points sampled from the reference frame and tracking points with the distorted optical flow (PRTD), as well as points sampled from the reference frame and tracking points with the reference optical flow (PRTR). These different process types are tested on the LIVE video database and the IVP database, and the experimental results are given in Fig. 10. As can be seen, the performance (SRCC) of each method does not vary much, and the fluctuation of SRCC is within 0.005 on the two databases, which suggests that, actually, which video frame points should be sampled from and which optical flow should be used to track points are not critical things, since the experimental results are almost the same. Even PRTD seems to perform a little better, as interpreted in Section II, in this work, the type of PDTD is used since it is believed to be more natural.

Then, a set of experiments are performed to determine the proper trajectory length $\mathcal{L}$. $\mathcal{L}$ is set from 12 to 27, and the results are shown in Fig. 11. Both the SRCC on the LIVE and the IVP reach the highest value when $\mathcal{L}$ is set to 18, and clearly, the two curves represents an apparent trend, increasing in the first stage, and then decreasing. Noted that when $\mathcal{L}$ is set to 27, SRCC on the IVP database increases a little bit, but drops more on the LIVE database. Besides, the fluctuation of SRCC on the IVP database is still very small (ranging from 0.9057 to 0.9090), and the one on the LIVE database

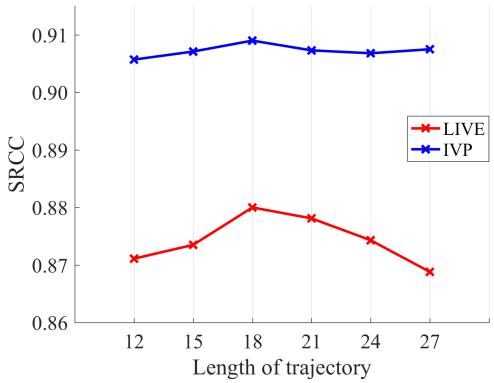


Fig. 11: Performance (SRCC) comparison of different length of trajectory on LIVE and IVP databases

appears (ranging from 0.8688 when $\mathcal{L} = 27$, to 0.8800 when $\mathcal{L} = 18$), which might indicate that the tracking length plays a more influential role on the LIVE database. Hence, the little bit increased on the IVP database when $\mathcal{L} = 27$ can be seen as a disturbance, but the dropping on the LIVE database is real, which further pushes us to believed that 18 frames is the proper trajectory length.

Next, efforts are also made to choose the local region size W. Similar to $\mathcal{L}$, different possible values are tested on the LIVE database and the IVP database, and the experimental results can be seen in Fig. 12. SRCC on the LIVE database increases at the beginning, reaches the highest value at $W = 48$, and decreases afterwards, while performance on the IVP database is always decreasing. It seems a little weird that these two databases present different trends, but in another point of view, it is duet to the difference that many VQA methods can not perform well on both of the two databases. And as a compromise, W is set to 48 in this work, which guarantees the best SRCC on the LIVE database, and a relatively good one on the IVP database.

Finally, different values are tried for $\mathcal{K}$, the histogram bins of temporal representation in Section III-A. Possible values are ranging from 4 to 16, and experimental results are shown in

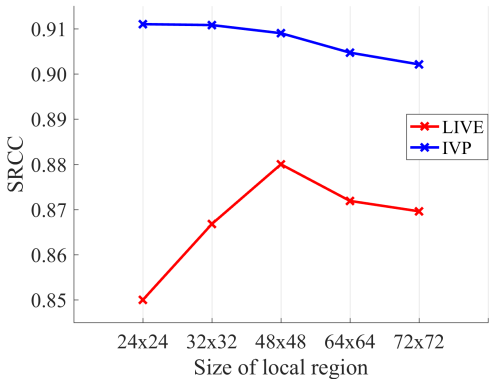


Fig. 12: Performance (SRCC) comparison of different size of local region on LIVE and IVP databases

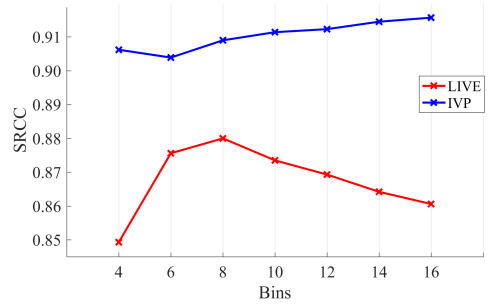


Fig. 13: Performance (SRCC) comparison of different bins for optical flow histogram on LIVE and IVP databases

Fig. 13. Since SRCC on the LIVE database achieve the best when $\mathcal{K} = 8$, while increases slowly on the IVP database, $\mathcal{K}$ is set to 8 in this work for a good trade-off between databases. Analogous to Fig. 12, the two databases present different preferences for parameters, which further shows the variation of VQA databases, and brings more challenges for a generalized VQA modeling.

V. Conclusion

In this paper, we have proposed a novel FR VQA model by measuring degradation along motion trajectory. Trajectories have been firstly generated based on point selection and position prediction. Then, motion velocity (magnitude and orientation) and motion content along trajectories have been extracted, and quality degradations on motion velocity and motion content have been measured (pooling into the temporal quality as well as the joint spatial-temporal quality). Finally, incorporating the spatial quality degradation, a novel trajectory-based model for FR VQA has been generated. Experimental results on five public VQA databases have demonstrated that our model achieves a remarkable improvement over existing VQA models, obtains a good generalization within different databases, possesses a nice robustness with distinct IQA metrics, and performs consistently with human perception. Moreover, the computational cost of our proposed model is also much lower than most existing relevant models.

References

[1] Q. Wu, H. Li, Z. Wang, F. Meng, B. Luo, W. Li, and K. N. Ngan, "Blind image quality assessment based on rank-order regularized regression," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2490–2504, Nov 2017.

[2] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[3] C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," *Optical Engineering*, vol. 42, no. 1, pp. 265–273, Jan. 2003.

[4] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, vol. 24, no. 12, pp. 61–69, Dec. 2007.

[5] S. Treue and J. H. R. Maunsell, "Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas," *Journal of Neuroscience*, vol. 19, no. 17, pp. 7591–7602, Sep. 1999.

[6] Y. Fang, W. Lin, B. Lee, C. T. Lau, Z. Chen, and C. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187–198, 2012.

[7] S. Treue and J. C. M. Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, no. 6736, pp. 575–579, Jun. 1999.

[8] W. Zhang and H. Liu, "Study of saliency in objective video quality assessment," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1275–1288, 2017.

[9] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[10] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, Apr. 2009.

[11] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, p. 013016, Feb. 2014.

[12] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 200–213, Jan. 2014.

[13] M. K. and S. S. Channappayya, "An Optical Flow-Based Full Reference Video Quality Assessment Algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[14] P. Peng, D. Liao, and Z.-N. Li, "An efficient temporal distortion measure of videos based on spacetime texture," *Pattern Recognition*, vol. 70, pp. 1–11, Oct. 2017.

[15] L. He, W. Lu, C. Jia, and L. Hao, "Video quality assessment by compact representation of energy in 3d-DCT domain," *Neurocomputing*, vol. 269, pp. 108–116, Dec. 2017.

[16] S. Treue and J. H. Maunsell, "Attentional modulation of visual motion processing in cortical areas MT and MST," *Nature*, vol. 382, no. 6591, pp. 539–541, Aug. 1996.

[17] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annual Review of Neuroscience*, vol. 28, pp. 157–189, 2005.

[18] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Research*, vol. 38, no. 5, pp. 743–761, Mar. 1998.

[19] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[20] J. Shi *et al.*, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[21] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 363–370.

[22] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 525–535, Jun. 2012.

[23] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[24] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1700–1705, Nov. 2013.

[25] L. Li, Y. Zhou, K. Gu, W. Lin, and S. Wang, "Quality assessment of dibr-synthesized images by measuring local geometric distortions and global sharpness," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 914–926, April 2018.

[26] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement," *IEEE Trans. Cybernetics*, vol. 46, no. 1, pp. 284–297, 2016.

[27] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[28] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[29] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," in *Human Vision and Electronic Imaging XVII*, vol. 8291. International Society for Optics and Photonics, 2012, p. 82911K.

[30] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[31] 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.

[32] F. Zhang, S. Li, L. Ma, Y. Wong, and K. Ngan, "Ivp subjective quality video database," *The Chinese University of Hong Kong, http://ivp. ee. cuhk. edu. hk/research/database/subjective*, 2011.

[33] S. Wolf and M. Pinson, "Video quality model for variable frame delay (vqm-vfd)," *US Dept. Commer., Nat. Telecommun. Inf. Admin., Boulder, CO, USA, Tech. Memo TM-11-482*, 2011.

[34] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2013.

[35] L. Xu, W. Lin, L. Ma, Y. Zhang, Y. Fang, K. N. Ngan, S. Li, and Y. Yan, "Free-energy principle inspired video quality metric and its use in video coding," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 590–602, 2016.

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

[37] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "Speed-qa: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, 2017.

[38] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. Ieee, 2003, pp. 1398–1402.

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

[40] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[41] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[42] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator," *IEEE Access*, vol. 4, pp. 5579–5590, 2016.

[43] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro-and macro-structures," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 3903–3912, 2017.