https://doi.org/10.2352/EI.2025.37.9.IQSP-251

Optimizing Frame Selection for Improved Video Quality Assessment Through Embedding Similarity

Abderrezzaq Sendjasni [1]***, Mohamed-Chaker Larabi*** [1] ***and Seif-Eddine Benkabou*** [2]
[1] ***CNRS, Xlim UMR 7252, Université de Poitiers, France***
[2] ***LIAS, Univ. Poitiers, France***

Abstract

This paper proposes a novel frame selection technique based on embedding similarity to optimize video quality assessment (VQA). By leveraging high-dimensional feature embeddings extracted from deep neural networks (ResNet-50, VGG-16, and CLIP), we introduce a similarity-preserving approach that prioritizes perceptually relevant frames while reducing redundancy. The proposed method is evaluated on two datasets, CVD2014 and KonViD-1k, demonstrating robust performance across synthetic and real-world distortions. Results show that the proposed approach outperforms state-of-the-art methods, particularly in handling diverse and in-the-wild video content, achieving robust performances on KonViD-1k. This work highlights the importance of embedding-driven frame selection in improving the accuracy and efficiency of VQA methods.

Keyworkds: *Video quality assessment, Frame selection, Embedding similarity, Deep Neural Networks.*

Introduction

The proliferation of video content across digital platforms, from ultra-high-definition streaming and immersive augmented reality/virtual reality experiences to real-time video conferencing and user-generated social media, has intensified the demand for robust video quality assessment (VQA) methodologies. As global internet video traffic is projected to dominate 82% of all consumer internet traffic by 2025 [17], the imperative to deliver visually coherent and perceptually satisfying content grows exponentially. VQA serves as the cornerstone for this endeavor, aiming to objectively evaluate perceptual quality by simulating human visual sensitivity to artifacts caused by compression, transmission errors, or computational bottlenecks [21, 23]. Its applications are critical: streaming platforms leverage VQA to optimize bitrate-ladder algorithms [5], telecom operators use it to monitor network-induced impairments, and developers rely on it to refine encoding pipelines, all while balancing bandwidth constraints against the viewer's quality-of-experience (QoE).

Conventional VQA approaches, such as PSNR, SSIM [20], and VMAF [10], are often used at the frame level. However, they depend on uniform or random frame sampling to reduce computational overhead. This strategy risks overlooking the temporal dynamics and spatial complexity inherent in video content. For instance, a high-motion sports sequence with rapid scene transitions or a slow-paced cinematic shot with nuanced textures may be inadequately represented by sparse, arbitrarily selected frames. Human observers, inherently attuned to motion saliency and spatial detail, perceive quality degradation unevenly across such varied content, yet, it is poorly captured by traditional sampling. Studies indicate that inconsistent frame selection can skew vision tasks in critical edge cases, leading to degraded performances [14]. This gap underscores the need for a frame selection paradigm that prioritizes perceptually critical frames, such as those encapsulating salient motion, structural complexity, or scene transitions, to align algorithmic assessments with human judgment.

The emergence of learning-based models addressed these limitations through adaptive sampling. For instance, TLVQM [8] employed a two-level approach, combining short-term frame-level features, such as blur and noise, with long-term video-level features, such as scene cuts, effectively balancing dense sampling of low-level details with sparse high-level semantic analysis. VSFA [9] introduced saliency-guided temporal pooling, weighting frames by visual importance using pre-trained saliency maps, though it retained uniform sampling upfront. In contrast, GST-VQA [1] adopted a dual-branch spatiotemporal sampler, extracting global motion patterns via 3D-CNNs on downsampled clips while preserving local textures through 2D-CNNs on high-resolution patches, a hybrid approach to capture multi-scale distortions. CoINVQ [19] leveraged contrastive snippet sampling, selecting video segments with similar/dissimilar distortions for self-supervised pretraining, bypassing the need for reference videos. Meanwhile, transformer-based models like Full-resolution Swin-T [11] used windowed attention across non-overlapping frame patches, enabling full-resolution processing without downsampling but at high computational cost. Addressing efficiency, FAST-VQA [22] pioneered fragmented grid sampling, splitting videos into short clips and spatially subsampling mini-patches, such as 32×32 grids, to drastically reduce input size while retaining global quality perception—achieving 100× speed gains over conventional methods. These advancements highlight a shift from naive uniform sampling to content-aware, efficient spatiotemporal sampling, balancing perceptual accuracy with computational feasibility, yet challenges remain in handling ultra-long videos and preserving fine-grained details.

Emerging advancements in representation learning offer a promising direction toward achieving an optimized frame selection for VQA, embedding similarity [18, 6]. By quantifying semantic and structural relationships between frames in a high-dimensional feature space, this approach enables the identification of key frames that collectively represent a video's perceptual essence. Unlike naive sampling, embedding-driven selection adapts to content characteristics, ensuring that frames with high informational entropy, such as dynamic movements and texture-rich scenes, are prioritized. Such a method not only mirrors the human visual system's adaptive attention but also bridges the gap between algorithmic efficiency and assessment accuracy. By do-

ing so, a dual advantage critical for next-generation VQA systems operating at scale is achieved. This paper explores how optimizing frame selection through embedding similarity can redefine video quality evaluation, offering a pathway to assessments that are both computationally efficient and perceptually faithful. To this end, we introduce a frame selection through embedding similarity analysis. This ensures that perceptually critical frames are retained to derive quality evaluation. The key contributions are:

- We introduce an embedding similarity-based selection method using the similarity-preserving technique to prioritize frames with high perceptual relevance.
- We conduct a comprehensive evaluation using embeddings from ResNet-50 [3] (for texture and shape features), VGG-16 [16] (for spatial details sensitivity), and CLIP [13] (for semantic context awareness). This ensures capturing diverse distortion characteristics and analyzing the resilience of the proposed selection method to various embeddings.
- We validate on two datasets, CVD2014 (with synthetic distortion) and KonViD-1k (with authentic user-generated content distortions), to demonstrate the robustness across synthetic and real-world scenarios.

In the following section, we present the proposed methodology for improving VQA through efficient frame selection.

Methodology

In this section, we present the proposed framework for optimizing VQA through embedding similarity-based frame selection. The framework consists of three key stages. First, a preprocessing and embedding generation step extracts feature representations from video frames using a deep neural network-based encoder. Second, an embedding similarity-based selection mechanism identifies the most representative frames by analyzing feature space similarities, thereby reducing redundancy while preserving essential visual information. Finally, a quality estimation stage evaluates video quality based on the selected frames and their corresponding patches.

Preprocessing

The preprocessing stage consists of two fundamental steps: patch extraction and feature encoding. This stage ensures that video frames are effectively represented in a compact feature space while retaining essential visual information for subsequent processing.

Given an input video $\mathscr{V} = \{t_1, t_2, \ldots, t_n\}$, where each frame $t_k \in \mathbb{R}^{H\times W\times 3}$ has spatial dimensions $H \times W$ and three color channels, patches are extracted using a uniform sampling strategy. Let $P_k = \{p_{k,1}, p_{k,2}, \ldots, p_{k,N_p}\}$ denote the set of patches sampled from frame t_k, where each patch $p_{k,i}$ has a fixed resolution of 224×224 pixels. In this framework, $N_p = 5$ patches per frame are selected to ensure diverse spatial coverage while maintaining computational efficiency. The uniform sampling approach ensures that patches are distributed evenly across each frame, capturing both background and foreground details.

Each extracted patch $p_{k,i}$ is subsequently processed by a visual feature encoder $\mathscr{E} : \mathbb{R}^{224\times 224\times C} \to \mathbb{R}^d$, which maps the patch into a d-dimensional feature embedding:

$$\mathbf{e}_{k,i} = \mathscr{E}(p_{k,i}), \quad \mathbf{e}_{k,i} \in \mathbb{R}^d. \tag{1}$$

Here, $\mathbf{e}_{k,i}$ represents the feature embedding corresponding to the i-th patch from frame t_k. The encoder $\mathscr{E}$ is implemented as a deep neural network trained to extract discriminative features related to texture, structure, and spatial coherence. In this study, we employ three widely used visual encoders, including ResNet-50 [3], VGG-16 [16], and CLIP-B/32 [13], to generate robust and diverse feature embeddings. These models have demonstrated strong performance in various vision tasks, making them suitable for extracting meaningful representations from video content.

The final output of this stage is a set of embeddings representing all extracted patches across the entire video. Formally, the complete embedding representation of the video $\mathscr{V}$ is given by:

$$\mathbb{E}(\mathscr{V}) = \bigcup_{k=1}^{m} \bigcup_{i=1}^{N_p} \mathbf{e}_{k,i}, \quad \mathbf{e}_{k,i} = \mathscr{E}(p_{k,i}), \tag{2}$$

where $\mathbb{E}(\mathscr{V})$ denotes the set of all feature embeddings extracted from the video, m is the total number of frames, and N_p is the number of patches per frame. These embeddings serve as the foundation for the subsequent frame selection stage, where redundant information is minimized while preserving the most relevant visual features for quality assessment.

Embedding similarity-based selection

The objective of this stage is to identify the most relevant embeddings while discarding those that do not contribute meaningful information for quality assessment. To achieve this, we leverage similarity preservation, a widely used concept in feature selection for high-dimensional data [24], and integrate it with residual analysis to refine patch selection.

Selecting an optimal subset of features using the similarity-preserving approach involves learning a transformation matrix $\mathbf{W}_j$ that maps the original high-dimensional embedding space $\mathbf{E}_j$ into a lower-dimensional representation $\mathbf{E}_j\mathbf{W}_j$, while preserving the similarity structure among the embeddings. This optimization problem is formulated as:

$$\min_{\mathbf{W}_j} \left\| (\mathbf{E}_j\mathbf{W}_j)(\mathbf{E}_j\mathbf{W}_j)^\top - \mathbf{S}_j \right\|_F^2 + \alpha \|\mathbf{W}_j\|_{2,1}, \tag{3}$$

where:

- $\mathbf{S}_j$ represents the similarity matrix computed in the original embedding space $\mathbf{E}_j$, using the Euclidean distance as the similarity measure.
- $\mathbf{W}_j \in \mathbb{R}^{d_j\times h}$ is the transformation matrix, where h ($h \ll d_j$) denotes the dimensionality of the reduced space $\mathbf{E}_j\mathbf{W}_j$.
- The term $\alpha\|\mathbf{W}_j\|_{2,1}$ enforces sparsity in $\mathbf{W}_j$, promoting the selection of only the most relevant features.

This formulation ensures that the transformed embeddings retain the structural relationships present in the original feature space while reducing redundancy.

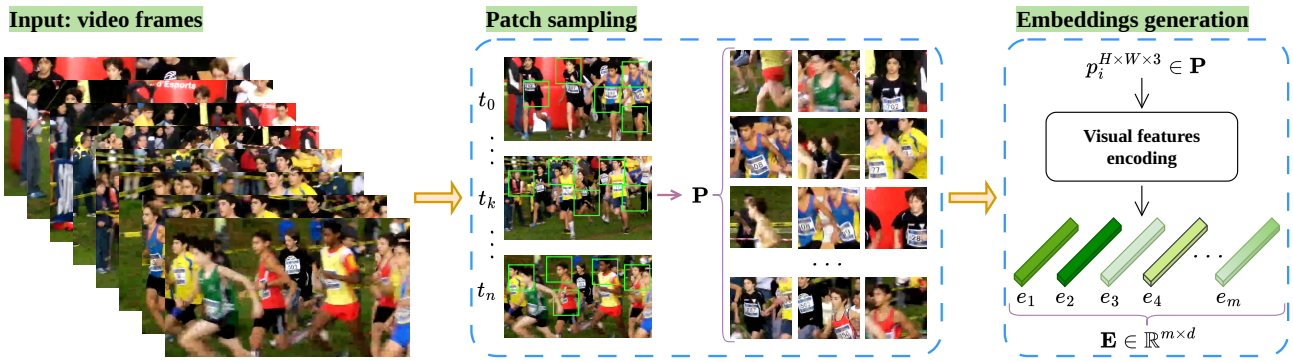


Figure 1: Workflow for extracting visual embeddings from video frames. Patches are uniformly sampled from the input video frames, and each visual patch $p_i \in \mathbf{P}$ is represented by an embedding $\mathbf{e}_i$.

To further refine the embeddings, we introduce residual analysis, based on the assumption that after projection into the reduced space, the similarity structure of relevant embeddings will be preserved, while that of irrelevant embeddings will not. To formalize this, we define a residual matrix $\mathbf{R}_i$, which quantifies the deviation of the transformed embeddings from the eigen-decomposition of the similarity matrix:

$$\mathbf{R}_i = (\mathbf{E}_i\mathbf{W}_i)^\top - \mathbf{Z}_i^\top - \Theta, \tag{4}$$

where:

- $\mathbf{Z}_i \in \mathbb{R}^{m\times h}$ is the eigen-decomposition of the similarity matrix, such that $\mathbf{S}_i = \mathbf{Z}_i\mathbf{Z}_i^\top$.
- Θ is a random matrix assumed to follow a multivariate normal distribution [15].

Each column of $\mathbf{R}_i$ corresponds to a patch $\mathbf{p}_i$ in the embedding space $\mathbb{E}(\mathscr{V})$. A large ℓ_2-norm of $\mathbf{R}_i(:,i)$ indicates that the patch $\mathbf{p}_i$ is likely irrelevant, as its similarity structure is not well preserved.

Using this residual-based filtering mechanism, we define the final patch selection optimization problem as follows:

$$\min_{\mathbf{W}_i,\mathbf{R}_i} \|\mathbf{E}_i\mathbf{W}_i - \mathbf{Z}_i - \mathbf{R}_i^\top\|_F^2 + \alpha\|\mathbf{W}_i\|_{2,1} + \beta\|\mathbf{R}_i\|_{2,1}, \tag{5}$$

where β is a regularization hyperparameter that controls the sparsity of $\mathbf{R}_i$ and, consequently, the number of patches retained. By employing this embedding similarity-based selection approach, we effectively reduce the number of patches required for accurate quality prediction. This selective filtering allows us to retain the most informative embeddings while discarding redundant or irrelevant ones.

In the next stage, these selected embeddings serve as input for a regression model, which learns a mapping between the retained features and the mean opinion scores (MOS). The details of this quality estimation process are discussed in the following section.

Quality estimation

In this stage, the goal is to predict the perceptual quality of the video content by mapping the selected visual embeddings to quality scores. The selected embeddings, denoted as $\mathbb{E}(\mathscr{V})_{\text{sel}} \subset \mathbb{E}(\mathscr{V})$, encapsulate the most informative features for quality prediction after the embedding selection process.

We formulate the quality assessment task as a regression problem. Let $\mathbf{e}_i \in \mathbb{R}^h$ represent the i-th selected embedding and $y_i \in \mathbb{R}$ denote the corresponding Mean Opinion Score (MOS) that reflects the subjective quality of the video. The regression model $f : \mathbb{R}^h \to \mathbb{R}$ is designed to learn the mapping between the feature space and the MOS. In its simplest linear form, the model can be expressed as:

$$f(\mathbf{e}_i) = \mathbf{w}^\top\mathbf{e}_i + b, \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^h$ is the weight vector and $b \in \mathbb{R}$ is the bias term. To capture potential non-linear relationships between the embeddings and the quality scores, more complex models such as a multi-layer perceptron (MLP) may be employed. In that case, the model is formulated as:

$$f(\mathbf{e}_i) = g(\mathbf{e}_i;\Theta), \tag{7}$$

where g represents the non-linear function parameterized by $Thet$ (comprising weights and biases across multiple layers).

To train the regression model, we minimize a loss function that measures the discrepancy between the predicted quality scores and the ground truth MOS. To this end, we use the mean squared error (MSE) loss, given by:

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{e}_i))^2, \tag{8}$$

where N is the number of training samples. Once the model is trained, it predicts the quality score for any new input embedding $\mathbf{e}_i$. For a given video, quality assessment can be performed by aggregating the predictions across the selected patches. The overall video quality score Q is then computed as an aggregation function using an average pooling:

$$Q = Avg\left(\{f(\mathbf{e}_i)\}_{i=1}^{N_{\text{sel}}}\right), \tag{9}$$

where N_{sel} is the total number of selected embeddings.

Results and discussion

Experimental details

Implementation: the proposed study is implemented using PyTorch. All experiments are conducted on a server equipped with an Intel Xeon Silver 4208 2.1GHz CPU, 192GB of RAM, and an Nvidia Tesla V100S GPU with 32GB of memory. We train the regression model for 200k iterations with a batch size of 128. We employ the Adam optimizer [7] to update the parameters of the model, with an initial learning rate $1e-4$.

Datasets: We used two benchmark dataset, the CVD2014 [12] and KonViD-1k [4]. The CVD2014 dataset comprises 234 videos with synthetic distortions, including compression artifacts, Gaussian noise, and blurring, derived from 12 reference videos. It is annotated with MOS and is widely used for evaluating VQA algorithms under controlled conditions. In contrast, the KonViD-1k dataset contains 1,2k authentic, user-generated videos with naturally occurring distortions, such as compression artifacts, camera noise, and motion blur. Annotated with MOS through large-scale subjective testing, KonViD-1k is designed for no-reference VQA in real-world scenarios. Together, these datasets provide a comprehensive framework for evaluating VQA methods, with CVD2014 focusing on specific synthetic distortions and KonViD-1k addressing the challenges of diverse, in-the-wild video content.

Evaluation criteria: we evaluate the performance using two metrics: Pearson linear correlation coefficient (PLCC) and Spearman rank correlation coefficient (SRCC). To account for scale discrepancies between predicted quality scores and subjective ratings, a non-linear mapping is applied to the predicted scores using a five-parameter logistic function [2] prior to calculating the performance metrics.

In addition to PLCC and SRCC, we compute the relative contrast to mesure how much the performance obtained using the selection algorithm deviates from the baseline as the selection rate varies. It helps in understanding the trade-offs between reatined embeddings rate and model performance. The relative contrast is a metric used to quantify the difference between a model's performance at a given selection rate and its baseline performance. It is expressed as a percentage and calculated using the following formula:

$$\text{Relative contrast} = \frac{\text{Perf}_{\texttt{rate}} - \text{Perf}_{\texttt{base}}}{\text{Perf}_{\texttt{base}}} \times 100 \qquad (10)$$

where $\text{Perf}_{\texttt{rate}}$ is the performance metric (PLCC or SRCC) at a specific selection rate and $\text{Perf}_{\texttt{base}}$ is the baseline performance metric (PLCC or SRCC) achieved with the full training set without selection.

Performance analysis

With the intent to assess the effictiveness of the proposed selection algorithm, we first analyze the performance without selection considered as the baseline to compare with. The baseline performances are summerized in Table 1. Based on the obtained performances on the CVD2014 and KonViD-1k datasets, we can see that CLIP-B/32 consistently achieves the highest correlation with human subjective scores, outperforming ResNet-50 and VGG-16 in both PLCC and SRCC. On CVD2014, CLIP-B/32 attains a PLCC of 0.841 ($\pm$0.020) and an SRCC of 0.830 ($\pm$0.015), while on KonViD-1k, it achieves a PLCC of 0.830 ($\pm$0.015) and an SRCC of 0.822 ($\pm$0.018). This superior performance is attributed to CLIP's ability to encode semantic context and cross-modal relationships, which are critical for capturing perceptual quality across diverse video content. In contrast, ResNet-50 and VGG-16, while competitive, exhibit slightly lower performance, likely due to their reliance on spatial and textural features without the semantic depth of CLIP. These findings highlight the importance of semantically rich embeddings for VQA and establish a robust baseline for evaluating the effectiveness of frame selection strategies in improving video quality.

Table 1: Baseline performances without selection on CVD2014 and KonViD-1k datasets. The best results in each column are highlighted in **bold red**.

Encoder	PLCC ($\pm$ Std)	SRCC ($\pm$ Std)
CVD2014		
ResNet-50	0.832 ($\pm$ 0.027)	0.820 ($\pm$ 0.018)
VGG-16	0.825 ($\pm$ 0.025)	0.818 ($\pm$ 0.017)
CLIP-B/32	**0.841** ($\pm$ 0.020)	**0.830** ($\pm$ 0.015)
KonViD-1k		
ResNet-50	0.820 ($\pm$ 0.020)	0.815 ($\pm$ 0.020)
VGG-16	0.815 ($\pm$ 0.020)	0.805 ($\pm$ 0.020)
CLIP-B/32	**0.830** ($\pm$ 0.015)	**0.822** ($\pm$ 0.018)

To thoroughly evaluate the impact of the proposed selection method on VQA performance, we analyze results across selection rates ranging from 10% to 90%. These results are compared to the baseline scenario, where no selection is applied (i.e., using 100% of the sampled patches). The performance trends in terms of PLCC and SRCC are illustrated in Fig.2 for CVD2014 and Fig.3 for KonViD-1k.

The curves depict the variations in PLCC and SRCC as a function of the selection rate, along with the relative contrast. As observed, the performance trends across both datasets highlight the effectiveness of the proposed embedding-based frame selection in enhancing VQA accuracy compared to the baseline. Notably, both PLCC and SRCC values increase until they reach the baseline performance at selection rates of 40–50% for CVD2014 and 60–70% for KonViD-1k, as indicated by the relative contrast curves. This demonstrates that reducing noisy frames and retaining more informative ones improve alignment with human subjective scores. The performance continues to improve with increasing selection rates, eventually surpassing the baseline, confirming the benefits of the proposed selection approach.

A performance plateau is observed for both datasets, but it occurs earlier on CVD2014 and later on KonViD-1k. This discrepancy can be attributed to the differences in content diversity and quality variations between the datasets. CVD2014 contains professionally captured content with relatively consistent distortions, allowing the model to reach optimal performance with a lower selection rate. In contrast, KonViD-1k consists of in-the-wild user-generated videos, which exhibit higher variability in content and distortions. As a result, a higher selection rate is required to capture the broader range of quality-related information, leading to a delayed performance plateau.

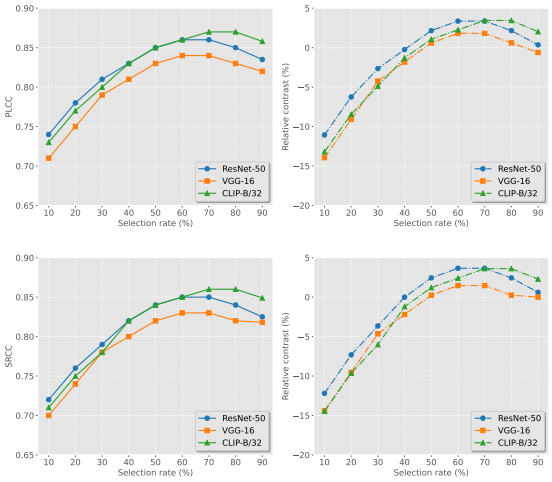


Figure 2: Impact of the selection rate on the performances and relative contrast to baseline versions on the CVD2014 dataset. (Left) Performance of PLCC and SRCC. (Right) Relative contrast compared to baseline performance, expressed as a percentage change.

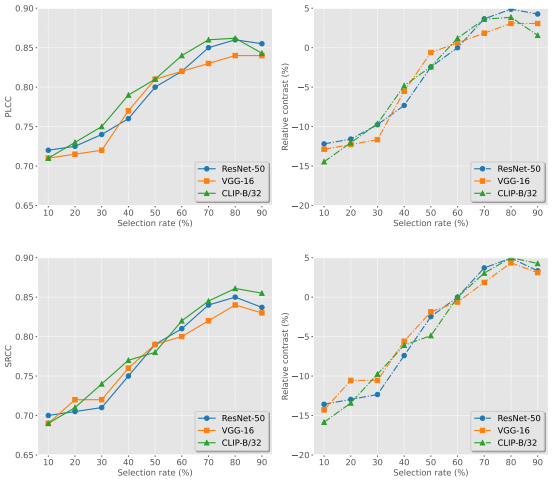


Figure 3: Impact of the selection rate on the performances and relative contrast to baseline versions on the KonViD-1k dataset. (Left) Performance of PLCC and SRCC. (Right) Relative contrast compared to baseline performance, expressed as a percentage change.

Performance comparaison

Table 2: Performance comparison with state-of-the-art methods on CVD2014 and KonViD-1k datasets. The best and second-best performances are respectively highlighted in **bold red** and **bold blue**.

Method	CVD2014		KonViD-1k	
	PLCC	SRCC	PLCC	SRCC
TLVQM [TIP, 2019]	0.850	0.830	0.724	0.732
VSFA [MM, 2019]	0.870	**0.868**	0.794	0.784
GST-VQA [TCSVT, 2022]	0.845	0.831	0.825	0.814
CoINVQ [CVPR, 2021]	0.844	0.830	0.764	0.767
Fr-Swin-T [ECCV, 2022]	**0.871**	0.868	0.838	0.841
FAST-VQA [ECCV, 2022]	**0.892**	**0.877**	0.850	**0.854**
Proposed (ResNet-50)	0.860	0.851	**0.855**	0.849
Proposed (VGG-16)	0.842	0.830	0.840	0.840
Proposed (CLIP-B/32)	0.869	0.866	**0.857**	**0.856**

To validate the proposed method against state-of-the-art approaches, we compared its performance using different visual encoders (ResNet-50, VGG-16, and CLIP-B/32) with several established methods, including TLVQM [8], VSFA [9], GST-VQA [1], CoINVQ [19], Fr-Swin-T [11], and FAST-VQA [22]. The results, summarized in Table 2, demonstrate that FAST-VQA achieves the highest performance on the CVD2014 dataset, with a PLCC of 0.892 and an SRCC of 0.877, indicating its strong alignment with subjective quality scores. However, on the KonViD-1k dataset, the proposed method with CLIP-B/32 embeddings outperforms all existing approaches, achieving the highest PLCC (0.857) and SRCC (0.856). This underscores the effectiveness of the embedding similarity-based frame selection in enhancing quality assessment, particularly for diverse and in-the-wild video content. Additionally, the proposed method with ResNet-50 embeddings achieves the second-best PLCC (0.855) on KonViD-1k, further validating its reliability. While Fr-Swin-T performs competitively on both datasets, it does not surpass the proposed method on KonViD-1k, suggesting that the proposed selection mechanism refines feature representations more effectively for real-world videos. Consistent performance is also observed with VGG-16 embeddings, highlighting the adaptability of the proposed framework across varying video content. These results emphasize the critical role of frame selection in improving video quality prediction, particularly in scenarios with diverse content, where traditional and advanced learning-based VQA methods often face challenges.

Conclusion

In this paper, we presented a novel approach to optimizing video quality assessment (VQA) through embedding similarity-based frame selection. By leveraging deep neural network embeddings (ResNet-50, VGG-16, and CLIP), we introduced a similarity-preserving technique that effectively identifies and retains perceptually relevant frames while discarding redundant or noisy ones. This approach addresses the limitations of traditional uniform or random sampling methods, which often fail to capture the temporal and spatial nuances critical to human perception.

Our experiments on the CVD2014 and KonViD-1k datasets demonstrated the robustness and effectiveness of the proposed method. The results showed that the embedding similarity-based selection significantly improves VQA performance, particularly in handling diverse and real-world video content. On KonViD-1k, the proposed method achieved state-of-the-art performance, outperforming existing approaches in terms of both PLCC and SRCC. Furthermore, the method demonstrated consistent performance across different encoders, with CLIP-B/32 embeddings yielding the best results due to their semantic richness and cross-modal capabilities.

Acknowledgments

This work is supported by The French Research Funding Agency (ANR) under project IMPROVED ANR-22-CE39-0006.

References

[1] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1903–1916, 2021.

[2] Video Quality Experts Group et al. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. *VQEG*, 2003.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017.

[5] Reza Kalan and Ismail Dulger. A survey on qoe management schemes for http adaptive video streaming: Challenges, solutions, and opportunities. *IEEE Access*, 2024.

[6] Zhao Kang, Yiwei Lu, Yuanzhang Su, Changsheng Li, and Zenglin Xu. Similarity learning via kernel preserving embedding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4057–4064, 2019.

[7] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.

[9] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2351–2359, 2019.

[10] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. Vmaf: The journey continues. *Netflix Technology Blog*, 25(1), 2018.

[11] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[12] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[14] Jian Ren, Xiaohui Shen, Zhe Lin, and Radomir Mech. Best frame selection in a short video. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 3212–3221, 2020.

[15] Y. She and AB. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] Cisco Team. Cisco annual internet report (2018–2023) white paper, 2020.

[18] Qitong Wang and Themis Palpanas. Deep learning embeddings for data series similarity search. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1708–1716, 2021.

[19] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13435–13444, 2021.

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[21] Stefan Winkler. *Digital video quality: vision models and metrics*. John Wiley & Sons, 2005.

[22] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pages 538–554. Springer, 2022.

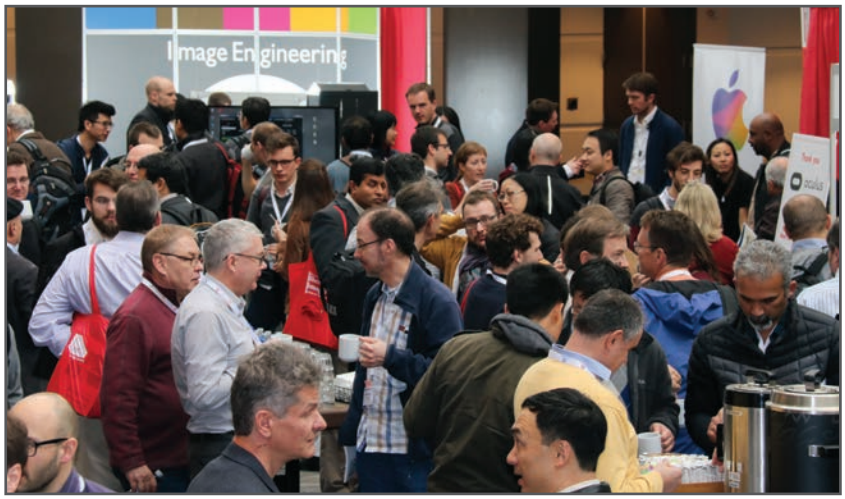
[23] Hong Ren Wu and Kamisetty Ramamohan Rao. *Digital video image quality and perceptual coding*. CRC press, 2017.

[24] Z. Zhao, L. Wang, H. Liu, and J. Ye. On similarity preserving feature selection. *IEEE TKDE*, 25(3):619–632, 2013.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org