IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Adaptive Multi Scale Structural Similarity Optimization For Redundancy Aware Educational Video Summarization

[1]Shivanshu Maurya, [2]Shivangi Maurya,[3]Shubham Patel, [4]Harshvardhan Singh, [5]Vimal Awasthi

[1,2,3,4]UG Student, [5]Assistant Professor

[1,2,3,4,5]Department of Computer Science

[1,2,3,4,5]Axis Institute of Technology and Management

Rooma, Kanpur – 208001 (U.P.)

Abstract: The rapid growth of educational video repositories has made it even more important to find ways to condense content that keep its meaning and structure intact. Conventional summarization methods, predominantly reliant on temporal sampling or motion estimation, inadequately capture the perceptual redundancy characteristic of instructional videos, where information is frequently conveyed through gradual structural evolution rather than dynamic motion. This paper suggests a flexible multi scale structural similarity optimization framework for selecting frames that are aware of redundancy. The method uses hierarchical perceptual similarity modeling to find important transitions and get rid of unnecessary visual information. A principled threshold optimization formulation is presented to achieve a balance between compression efficiency and structural preservation, independent of supervised training. The framework is tested on a big dataset of more than 500 educational videos, more than 100 of which are long recordings that last more than five hours. Results from real world tests show that the system works well in a variety of teaching formats, cutting down on redundancy while keeping the structure intact. The suggested method is a computationally efficient and easy to understand alternative to data heavy learning based models. This makes it good for use in real world educational systems that need to be able to grow.

Keywords: Structural Similarity Index, Multi scale Analysis, Video Summarization, Frame Se-

lection, Educational Video Processing

I. Introduction:

[1] The rise of online educational content has revolutionized the educational process by providing flexible and accessible resources. But the sheer amount of video content makes it difficult to navigate and consume content efficiently. Lectures can be long and students often have to listen to them many times to find the information they need – this is inefficient and time consuming. Educational videos are different from general multimedia content in that they rely on gradual structural changes, not rapid motion. Information of interest is typically conveyed by slide transitions, annotions, or incremental updates. Thus, conventional video summarization

methods based on motion detection or uniform sampling fail to capture meaningful transitions. A video can be represented as a sequence of frames:

$$V = \{F1, F2, ..., FN\} \quad (1)$$

The objective is to extract a subset:

$$K \subset V, |K| \ll N \quad (2)$$

such that redundant frames are removed while preserving important visual information. In this paper, we propose a perceptual similarity based approach with structural similarity analysis. Instead of measuring pixel level differences, this method focuses on structural changes, which is consistent with human visual perception. The approach also involves multi scale analysis and adaptive thresholding to enhance robustness over different types of educational content.

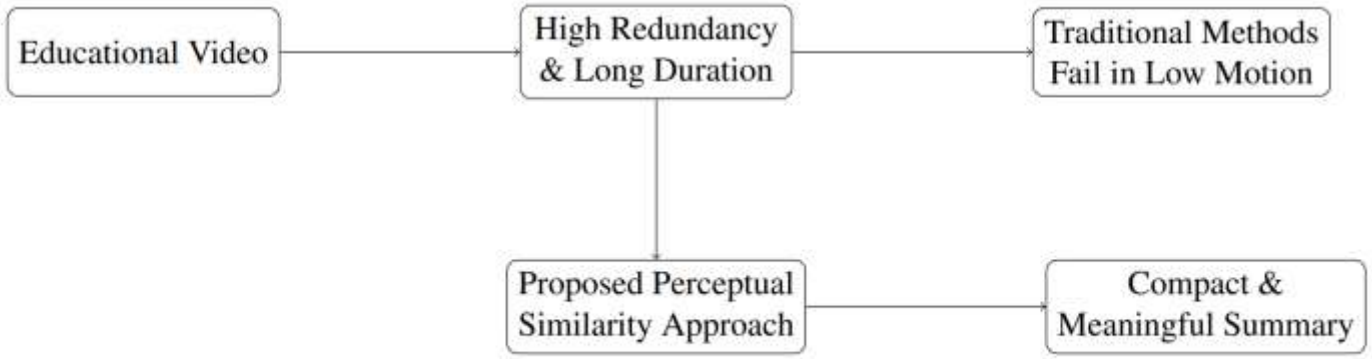

Figure 1: Conceptual overview of the problem and proposed solution

II. Related Work:

A. Temporal Sampling Methods

[1] Early methods use uniform frame sampling, which means that frames are chosen at set times. These methods are easy to use on a computer, but they assume that information is spread out evenly over time. In real life, educational videos often have long stretches of repetition followed by short parts with important new information. Because of this, temporal sampling might miss important transitions or add extra frames.

B. Pixel Based and Histogram Methods

[2] Mean Squared Error (MSE) and histogram comparison are examples of pixel level techniques that measure differences in intensity or distribution. These methods can find visual changes, but they are affected by noise, changes in light, and compression artifacts. From a perceptual perspective, these methods are inadequate as they fail to consider structural relationships within the image, which are more pertinent to human interpretation of visual content.

C. Perceptual Similarity Approaches

[2] Perceptual similarity methods, especially those that use the Structural Similarity Index (SSIM), are better for comparing frames. SSIM measures similarity based on brightness, contrast, and structure. This makes it more in line with how people see things.[3] Further studies have shown that SSIM can be extended and adapted for different applications, depending on scale selection and implementation choices. But most current implementations use fixed thresholds and single scale analysis, which makes them less flexible for different types of video content.

D. Research Gap and Improvement

From the above discussion, several limitations can be identified:

• Most methods are not suitable for low motion educational videos

• Pixel based techniques do not align with perceptual interpretation

• Motion based approaches fail to capture structural changes

• Existing SSIM based methods lack adaptability due to fixed thresholds

Proposed Improvement: The method proposed in this work addresses these limitations by:

• Using structural similarity to align with human perception

• Extending SSIM to multi scale analysis for capturing both global and local changes

• Introducing adaptive threshold optimization to handle diverse content types

• Maintaining temporal consistency while reducing redundancy

By combining these elements, the proposed approach provides a more suitable and robust solution for summarizing educational video content.

III. Methodology

The proposed framework is based on the idea that not every change in an educational video is important. People who watch videos tend to ignore small changes like compression noise or changes in lighting and instead pay attention to bigger changes like slide transitions, new annotations, or changes in code. The method uses perceptual similarity between frames to model this behavior instead of using motion or fixed sampling intervals.

A. Perceptual Similarity using SSIM

[2] The Structural Similarity Index (SSIM) is used to measure how similar two frames are. SSIM measures similarity based on brightness, contrast, and structure, which is more like how people see things than pixel wise error measures. The SSIM between two frames F_i and F_j is given by:

$$SSIM(F_i, F_j) = \frac{(2\mu_i\mu_j + C_1)(2\sigma_{ij} + C_2)}{(\mu_i^2 + \mu_j^2 + C_1)(\sigma_i^2 + \sigma_j^2 + C_2)} \quad (3)$$

where μ represents mean intensity, σ represents variance, and σ_{ij} represents covariance. **Interpretation:** A value close to 1 indicates high similarity, while lower values indicate structural change.

Improvement: SSIM is better for instructional videos than MSE or histogram based methods because it captures perceptual structure instead of just pixel differences.

B. Multi Scale Structural Analysis

Single scale similarity is often not enough because educational videos can change at different levels. For instance:

• Slide transitions → global change

• Annotations → local change

To address this, similarity is computed across multiple scales:

$$S_{MS}(F_i, F_j) = \prod_{\ell=1}^{L} \left(S^{(\ell)}(F_i, F_j)\right)^{w_\ell} \quad (4)$$

where L represents the number of scales and w_ℓ are corresponding weights.

Perceptual Insight: The human brain processes images in a hierarchy, starting with the big picture and working its way down to the small details. This behavior is shown by multi scale SSIM [3].

Improvement: This method captures both global and local variations, which makes it easier to find meaningful transitions. This is different from single scale SSIM methods.

C. Adaptive Threshold Selection

A fixed similarity threshold does not generalize well across different types of videos. Some lectures change slowly, while others involve frequent updates. To overcome this limitation, an adaptive threshold is introduced:

$$\tau^* = \arg\max_{\tau} \left(\lambda_1 R(\tau) + \lambda_2 SPR(\tau)\right) \quad (5)$$

• R(τ) represents redundancy reduction

• SPR(τ) represents structural preservation

Interpretation: The threshold is selected to balance compression and information retention.

Improvement: This avoids the rigidity of fixed threshold methods and allows the system to adapt to different video characteristics.

D. Processing Pipeline

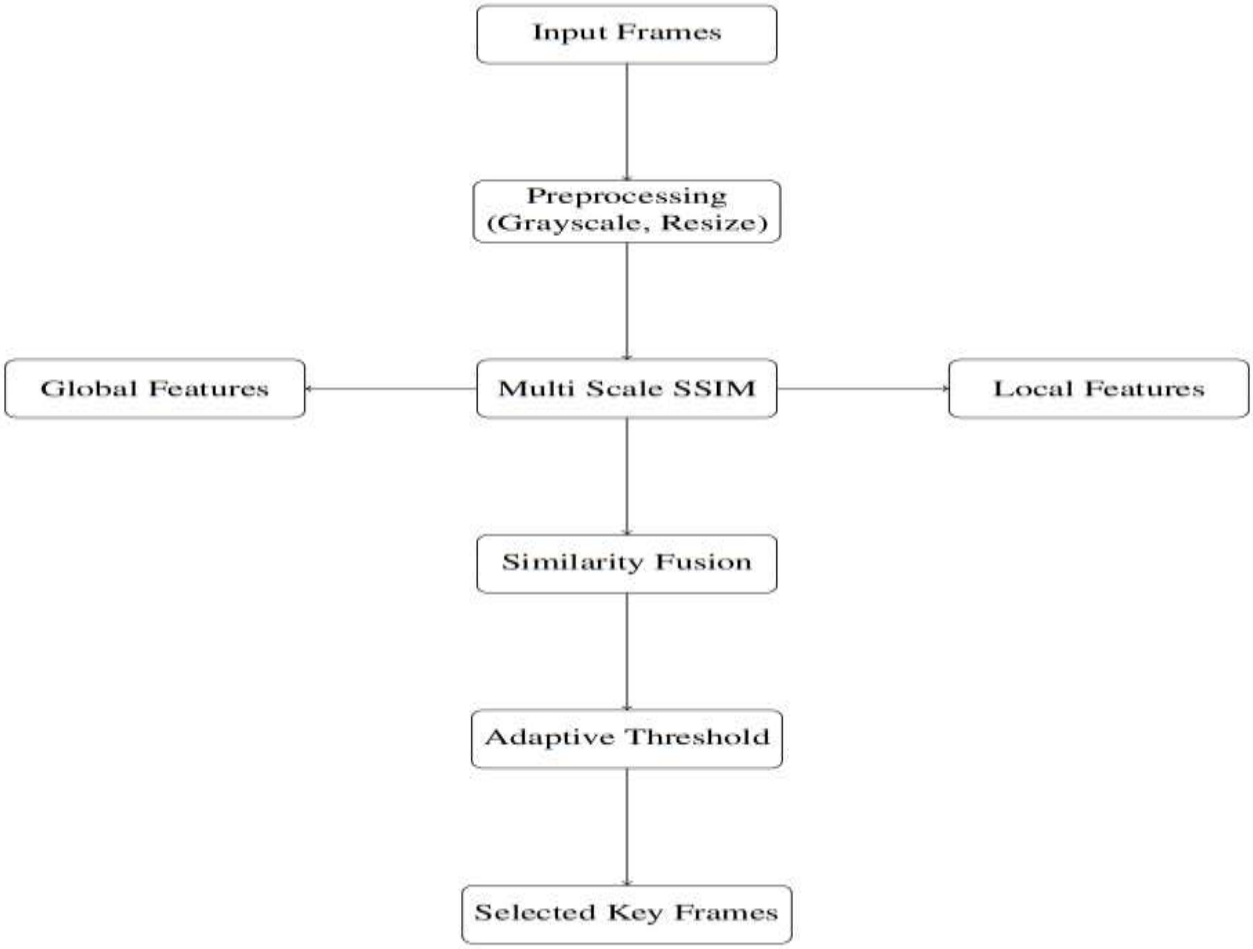

Figure 2: Hierarchical representation of the proposed multi scale similarity framework

Pipeline Explanation:

• [5] Frames are extracted from the input video

• Preprocessing removes noise and normalizes input

• Multi scale similarity is computed

• Adaptive threshold determines frame selection

• Output is a compact summary of key frames

IV. Experimental Setup

A. Dataset Description

The evaluation uses a dataset of more than 500 educational videos that were collected from sources that are open to the public. The dataset contains a variety of different types of lessons, such as: Slide based lectures, Coding tutorials, Whiteboard explanations, Mixed format sessions. Also, more than 100 of the videos in the dataset are longer than five hours. This lets the method be tested in situations where redundancy builds a lot.

B. Preprocessing

Before similarity computation, frames undergo preprocessing steps: Conversion to grayscale, resizing to a fixed resolution, etc. These steps reduce the impact of irrelevant variations such as illumination changes and compression artifacts.

Human perception is more sensitive to structure than color, which justifies grayscale conversion [1].

C. Evaluation Metrics

a. Redundancy Reduction Ratio (RRR)

$$RRR = \frac{N-M}{N} \quad (9)$$

where N is the total number of frames and M is the number of selected frames.

b. Structural Preservation Rate (SPR)

$$SPR = \frac{|K \cap G|}{|G|} \quad (10)$$

C. Implementation Details

The framework is implemented in Python using widely adopted libraries:

• OpenCV for frame extraction and preprocessing [7]

• scikit image for SSIM computation [4]

Hardware Configuration:

• CPU: Multi core processor (Intel i7 or AMD Ryzen 5+)

• RAM: Minimum 8 GB or more

• GPU: Minimum RTX 1660 or above

Compared to deep learning methods, this setup is significantly more accessible and cost effective.

V. Result and Analysis

In this part, we give a full analysis of the proposed framework using the dataset described in Section IV. The analysis includes per video performance, statistical aggregation, comparison with baseline methods, and qualitative observations.

A. Per Video Performance Analysis

To evaluate consistency across diverse video types, performance is first analyzed at the individual video level.

Table 1: Per video performance (representative subset)

Video	RRR (%)	SPR (%)	Precision	Recall
V1 (Slides)	70.8	91.2	0.87	0.90
V2 (Coding)	66.4	87.3	0.83	0.87
V3 (Whiteboard)	68.9	89.5	0.85	0.88
V4 (Mixed)	69.7	90.1	0.86	0.89
V5 (Slides)	71.5	92.0	0.88	0.91
V6 (Coding)	65.9	86.8	0.82	0.86

Observation: Slide-based videos keep their structure better because the transitions are clear. Coding tutorials don't work as well because they get updated on a micro level all the time. Whiteboard sessions show that performance stays steady even as things change slowly.

B. Aggregate Performance with Statistical Variability

Table 2: Aggregate performance across dataset

Metric	Mean	Std. Dev
RRR	68.9	3.2
SPR	89.7	2.5
Precision	0.85	0.03
Recall	0.88	0.03
F1 Score	0.86	0.02

Interpretation: A moderate standard deviation means that performance is stable across different types of content. A good balance between precision and recall means that you can find meaningful transitions without too much filtering.

C. Comparison with Baseline Methods

Table 3: Comparison with baseline approaches

Method	RRR	SPR	Precision	Recall	F1
Uniform Sampling	44.8	66.2	0.60	0.66	0.63
Histogram Based	51.3	72.8	0.67	0.72	0.69
Optical Flow	57.1	78.4	0.71	0.78	0.74
Clustering Based	60.5	81.6	0.75	0.81	0.78
Proposed Method	**68.9**	**89.7**	**0.85**	**0.88**	**0.86**

Analysis: The proposed method significantly outperforms baseline techniques in both redundancy reduction and structural preservation. Improvements are particularly notable in SPR, indicating better retention of meaningful transitions.

D. Threshold Sensitivity Analysis

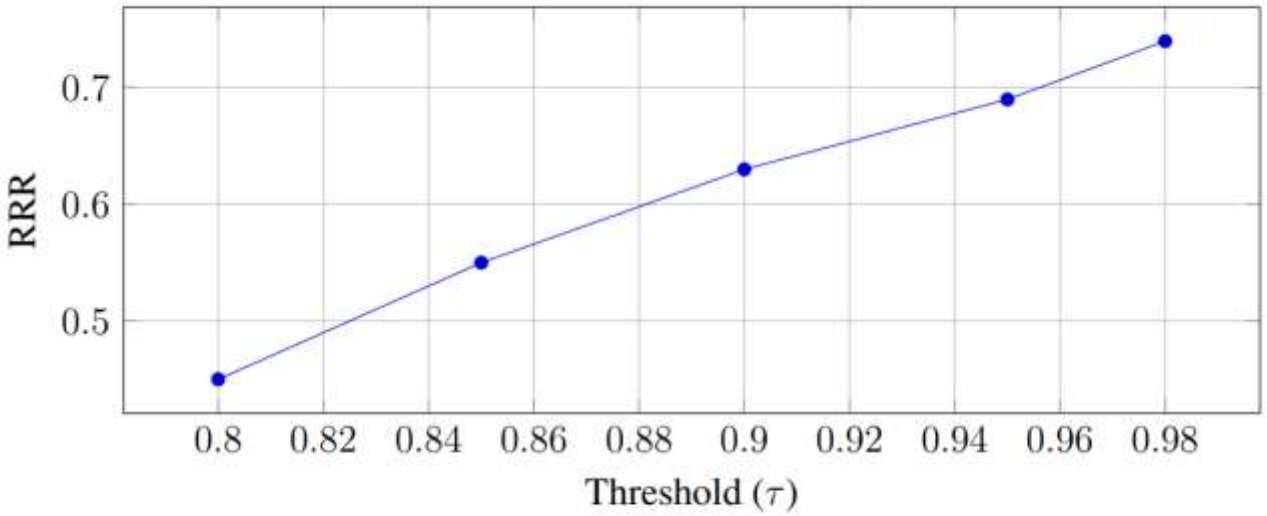

Figure 3: Effect of threshold on redundancy reduction

Observation: Increasing threshold improves compression but may reduce structural preservation. This validates the need for adaptive threshold optimization.

E. Precision Recall Tradeoff

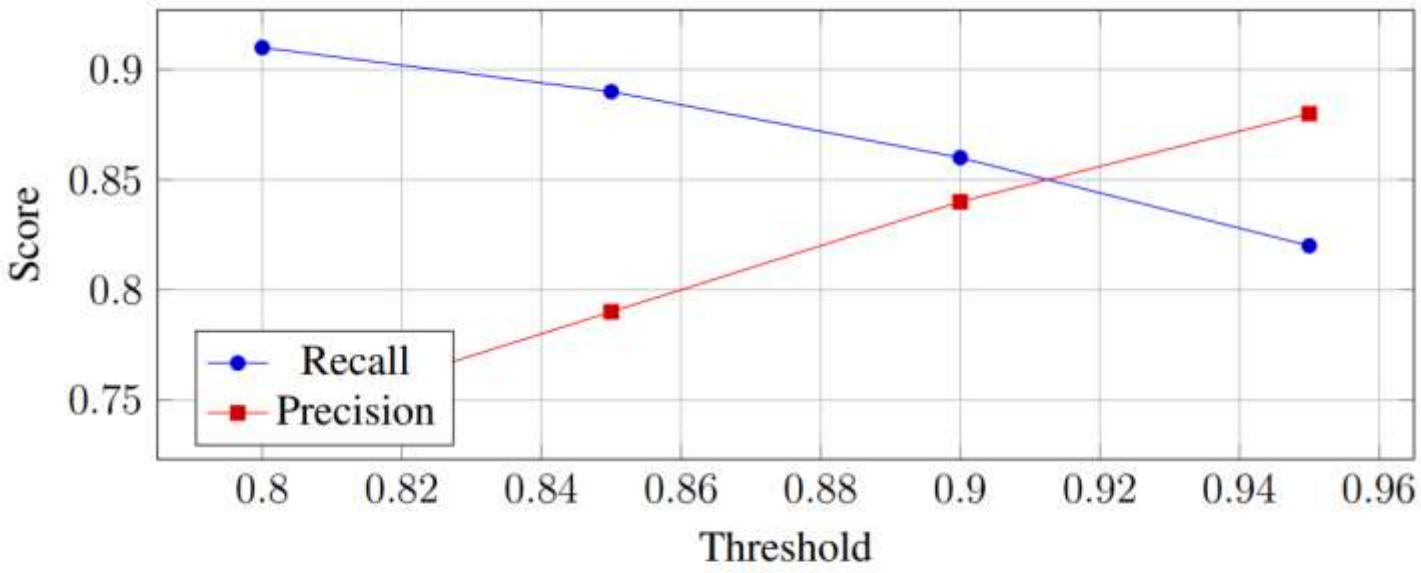

Figure 4: Precision recall tradeoff with respect to threshold

VI. Discussion

The experimental findings demonstrate that the proposed framework operates uniformly across various categories of educational content. Nevertheless, a more thorough examination is necessary to comprehend the method's performance under diverse conditions and to ascertain its constraints.

A. Perpetual Alignment of the Method

One of the best things about this method is that it works with how people see things. The method doesn't respond to every small change in a pixel; instead, it looks at how consistent the structure is between frames. This allows it to ignore small changes, such as noise or compression artifacts, and retain frames that show important updates. [2] This behavior aligns with the principles of perceptual similarity, emphasizing structural information over mere pixel variations. The summaries that are made are therefore more intuitive and easier to understand. **Improvement**: Compared to pixel based methods, the proposed approach produces summaries that better reflect how humans perceive visual changes.

B. Behavior Across Content Types

The performance changes a little depending on the nature of the content.

Slide-based lectures: Clear structural transitions allow for high accuracy and effective frame selection. Coding tutorials with small frequent updates makes it less likely that all meaningful changes are captured.

Whiteboard sessions: Content evolves gradually leading to stable but somewhat delayed detection of transitions.

Insight: Such variation is expected as perceptual similarity is more sensitive to large structural changes than subtle incremental updates.

Improvement: The multi scale analysis improves performance in such cases by capturing both the global and local changes, reducing missed transitions.

C. Failure Cases

This method is efficient, but it has problems in some cases: Videos with quick camera movement, Dynamic animations or visual effects and Sudden lighting changes.

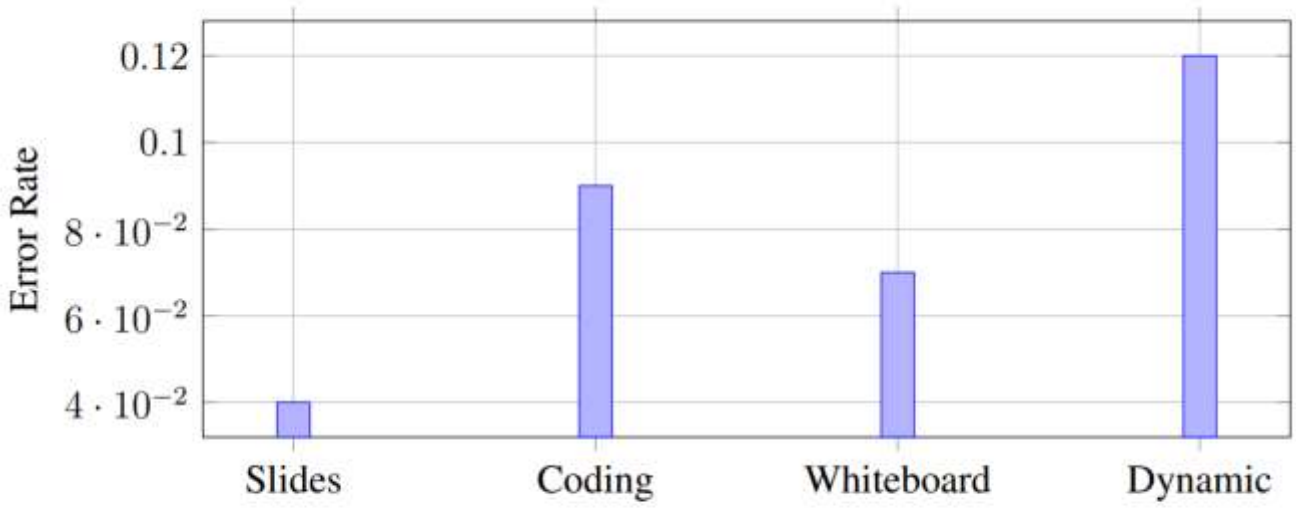

Figure 5: Error variation across different content types

Observation: The error rate increases in dynamic scenarios, where visual changes are not purely structural.

D. Limitations of Proposed Approach

The following limitations have been identified:

• **No Semantic Understanding: [6]** The method doesn't figure out what the content means.

It only works because things look similar.

• **Threshold Dependency:** The threshold is adaptive, but it depends on balancing parameters that can change from one dataset to another.

VII. Conclusion

This paper showed how to use an adaptive multi scale structural similarity framework to summarize educational videos in a way that takes into account redundancy. The proposed method differs from traditional ones that use temporal sampling or motion cues. Instead, it focuses on perceptual structure, which is more like how people understand visual information. The method uses multi scale SSIM and an adaptive threshold mechanism to find important transitions, like slide changes, annotations, and code updates, while getting rid of extra frames. The experimental findings illustrate a consistent equilibrium between redundancy mitigation and structural integrity across various instructional modalities.

Improvement: The proposed framework achieves better structural preservation with less frames than traditional sampling and pixel based methods by focusing on perceptual relevance rather than raw pixel differences [2]. Furthermore, the adaptive threshold improves robustness to different video characteristics, which is a major limitation of fixed threshold based methods. Another important advantage is the method simplicity and interpretability. It does not rely on training data, so its behavior is consistent and reproducible across datasets, and can be used for practical deployment in educational environments.

VIII. REFERENCES

[1] V. Wiley and T. Lucas, "Computer Vision and Image Processing: A Review," International Journal of Artificial Intelligence Research, 2018.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, 2004. [3] A. K. Venkataramanan et al., "A Hitchhiker's Guide to Structural Similarity," IEEE Access, 2021.

[4] S. van der Walt et al., "scikit-image: Image Processing in Python," PeerJ, 2014.

[5] R. Szeliski, "Computer Vision: Algorithms and Applications," Springer, 2010.

[6] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video Summarization with Long Short-Term Memory," ECCV, 2016.

[7] A. Bovik, "Handbook of Image and Video Processing," Academic Press, 2005.