International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# High-Performance Video Retrieval Using SIFT and Deep Learning Methods

**G. S. Naveen Kumar[1*], V. S. K. Reddy[2], Leela Kumari Balivada[3]**

**Submitted:** 02/10/2024 **Revised:** 15/11/2024 **Accepted:** 24/11/2024

**Abstract:** Shot boundary detection and key frame extraction are critical steps for video indexing, summarization, and retrieval. This paper proposes an advanced approach that combines a sophisticated SIFT (Scale-Invariant Feature Transform) keypoint matching algorithm with deep learning-based key frame extraction techniques. The SIFT-based method effectively captures both abrupt and gradual shot transitions, overcoming the limitations of existing algorithms that perform well only for abrupt changes. For key frame extraction, a deep learning framework leveraging convolutional neural networks (CNNs) for spatial feature representation and recurrent neural networks (RNNs/LSTMs) for temporal modeling is employed. This approach automatically identifies the most informative and representative frames while reducing redundancy, enabling efficient processing of large-scale video data. Extensive experiments on benchmark video datasets demonstrate that the proposed algorithms significantly outperform traditional methods in terms of accuracy, robustness, and computational efficiency, making them highly suitable for modern video analysis applications.

***Key words:*** *SIFT, Shot Boundary Detection, CNN, RNN, LSTM, Key Frame Extraction, CBVR*

## 1 Introduction

The exponential growth of video databases has been driven by the widespread availability of affordable video-capturing devices, decreasing storage costs, and high-speed broadband connectivity. This surge necessitates advanced and efficient video database systems for effective browsing, searching, and retrieval [1][2]. Traditional video retrieval methods rely on textual metadata, manual annotation, or tagging, which are time-consuming and require significant human effort. To address these limitations, Content-Based Video Retrieval (CBVR) systems have emerged as a more sophisticated alternative. Key prerequisites for CBVR include shot boundary detection and key frame extraction [3]. In this context, a shot is defined as a sequence of consecutive frames captured continuously between the start and stop operations of a camera. Shot boundaries, also known as shot transitions, separate consecutive shots and are typically classified into cut transitions and gradual transitions.

Shot boundary detection plays a crucial role in video indexing, classification, summarization, and retrieval, driving significant research advancements in this area. Several approaches have been proposed, including pixel-based, block-based, and histogram-based methods. In the pixel-based approach [4], individual pixels of consecutive frames are compared. While this method is computationally efficient and fast, it is highly sensitive to minor camera or object movements, which can falsely indicate dissimilarity between otherwise similar frames. Variations in illumination further affect its accuracy. The block-based method [5] addresses some of these limitations by dividing frames into smaller blocks and comparing corresponding blocks between frames. This approach is less sensitive than pixel-based techniques, but it is more computationally intensive and may struggle to detect rapidly moving objects. Histogram-based detection [6] relies on the statistical distribution of pixel intensities, making it robust to translation, rotation, and camera motion. However, it fails when two frames have different content but similar color distributions, leading to incorrect boundary detection.

[1]*Research Scholar, Department of ECE, JNTUK, Kakinada, India,*

[2]*Department of ECE, Malla Reddy University, Hyderabad, India.*

[3]*Department of ECE, UCEK, JNTUK, Kakinada, India.*

*gsrinivasanaveen@gmail.com *, vskreddy2003@gmail.com, leela8821@jntucek.ac.in*

In this paper, we propose a Scale-Invariant Feature Transform (SIFT)-based algorithm [7] integrated with deep learning techniques for robust shot boundary detection and key frame extraction. SIFT provides resilience to image rotation, scaling, illumination changes, and noise, enabling accurate detection of both abrupt and gradual shot transitions.

Following shot boundary detection, key frame extraction is performed using a deep learning framework. Specifically, Convolutional Neural Networks (CNNs) are employed to extract high-level spatial features from each frame, capturing semantic content, while Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks model temporal dependencies across frames to identify the most representative and informative frames within each shot.

This hybrid approach eliminates the need for manual or heuristic-based key frame selection, automatically reducing redundancy and retaining only the frames that best summarize the shot content. By combining SIFT for robust shot detection with CNN-RNN/LSTM networks for intelligent key frame extraction, the proposed method achieves high accuracy, efficiency, and scalability for video summarization, indexing, and retrieval tasks.

The organization of this paper is structured as follows: Section 1 presents the Introduction, highlighting the motivation and scope of the study. Section 2 details the proposed Shot Boundary Detection method using SIFT. Section 3 describes the Key Frame Extraction approach based on CNN and RNN/LSTM networks. Section 4 presents the Experimental Results, evaluating the performance of the proposed algorithm, and Section 5 concludes the paper, summarizing the findings and future directions [9].

## 2 Shot Boundary Detection

A shot is defined as a sequence of consecutive frames captured continuously between the start and stop of a camera. Consecutive shots are separated by a shot boundary or shot transition, which can be classified into cut transitions and gradual transitions. A cut represents an abrupt change between two shots, occurring only between the last frame of the first shot and the first frame of the next. In contrast, gradual transitions span multiple frames and are categorized into four types: fade in (transition from a constant image to a scene), fade out (transition from a scene to a constant image), dissolve (transition between two scenes through a combination of fade out and fade in), and wipe (transition between two scenes separated by a moving line). Accurate shot boundary detection requires effective feature extraction to evaluate the similarity or dissimilarity between frames. In this work, Scale-Invariant Feature Transform (SIFT) [10] is employed for robust feature extraction, providing resilience to scale, rotation, illumination changes, and noise.

### 2.1 SCALE INVARIANT FEATURE TRANSFORM:

Scale-Invariant Feature Transform (SIFT), developed by David G. Lowe, is a widely used algorithm for detecting and extracting local feature descriptors from an image or video frame. It is highly effective for object matching in large databases due to its robustness and invariance to changes in illumination, rotation, scaling, and added noise. The SIFT algorithm operates in four main stages: Detection of Scale-Space Extremes, Accurate Keypoint Localization, Orientation Assignment, and Descriptor Representation, enabling reliable and distinctive feature extraction for various computer vision applications

*1. Detection of Scale-Space Extreme:* In this stage, interest points or isolated points are detected in the scale space of an image. Isolated points are also called key points which become SIFT features. The key points are obtained by applying local extrema of difference of Gaussian (DoG) space to the frame. DoG scale space can be obtained from (1). The DoG image is represented as D(x, y, σ ) and can be computed from the difference of two nearby scaled images separated by a multiplicative factor k :

$$D(x,y,\sigma) = \big(G(x,y,k\sigma) - G(x,y,\sigma)\big) * I(x,y)$$

(1)

$$= L(x,y,k\sigma) - L(x,y,\sigma)$$

where L(x, y,σ ) is the scale space of an image,

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y)$$

(2)

G(x,y,σ) is a variable-scale Gaussian kernel defined as:

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

(3)

*2. Accurate Key Point Localization*

Candidate key point location and scale are determined in this stage. Key points are selected depending on the measure of stability. Accurate Key points are extracted by removing low contrast and noise affected candidate keypoints.

*3. Orientation Assignment*

In this stage, an orientation is assigned to each of the key points to construct the descriptor that is invariant to rotation. Orientation histogram of local gradients from the closest smoothed image L(x,y,σ) is used to calculate the orientation of a key point. For each image sample L(x, y) at this scale, the gradient magnitude m(x, y) and orientation θ (x, y) is computed using pixel differences:

$$m(x,y) = \sqrt{\big(L(x+1,y)-L(x-1,y)\big)^2+\big(L(x,y+1)-L(x,y-1)\big)^2} \quad (4)$$

$$\theta(x,y) = \tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \quad (5)$$

The orientation histogram covering the 360 degrees of major orientation with 36 bins

*4. Descriptor Representation*

A Descriptor is generated from local image gradients information for each key point. A 16×16 pixel region is used to calculate the feature vector descriptor with orientation histograms. Every 4×4 region is projected as a histogram with 8 bins and 8 directions. This process forms a SIFT feature vector descriptor with 128 directions shown in fig. 1.

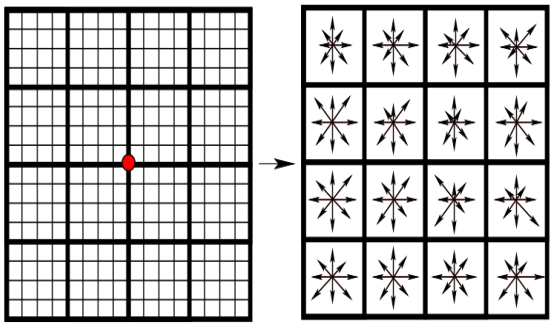

**Fig.1. Key point descriptor Generation with 128 directions**

## 2.2 Key Point Matching for Shot Boundary Detection

The key points in the current frame are compared with those of the next frame by using the k-Nearest Neighbor (kNN) Search. If the key point has minimum Euclidean distance, it is called the Nearest Neighbor. Nearest Neighbor Search is also termed as proximity search, the closest point search or the similarity search. k-Nearest Neighbor (kNN) Search algorithm is a simplest among the machine learning algorithm. k-Nearest Neighbor depends on Euclidean distance to determine whether the key points of the current frame are matched with those of the next frame.

Euclidean distance is a technique that measures the distance between two feature vectors. Let $\mathbf{x_i}$ be an input sample and $\mathbf{x_j}$ be testing sample with **p** number of feature (s=1, 2, …p). The Euclidean distance between sample $\mathbf{x_i}$ and $\mathbf{x_j}$ is defined as

$$d(x_i x_j) = \sqrt{\sum_{s=1}^{p}(x_{is}-x_{js})^2} \quad (6)$$

When the key points in the two frames are matched with each other, lines are drawn between the matching key points. If the distance between any two key points is less than 0.6, the matching is accepted. An example keypoint matching is shown in figure 2

If the number of matched keypoints between any two frames is greater than the threshold value, these two frames are related to the same region. Otherwise, shot boundary is detected.


**Fig.2. Keypoint Matching**

## 3 Key Frame Extraction

After successful shot boundary detection, the next critical step in video analysis is key frame extraction. Key frames are representative frames that summarize the content of a shot while eliminating redundancy, enabling efficient video indexing, retrieval, and summarization. Unlike traditional approaches that rely on handcrafted features such as color histograms, optical flow, or entropy-based metrics, the proposed method leverages deep learning techniques to automatically extract informative frames [11].

### 3.1 Deep Learning-Based Framework

The proposed framework integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to capture both spatial and temporal information:

Spatial Feature Extraction (CNN): Each frame within a shot is processed through a pre-trained CNN (e.g., ResNet, VGG) to extract high-level spatial features that capture semantic content such as objects, scenes, and textures. CNNs are highly effective in identifying distinctive frame-level characteristics that are crucial for selecting representative frames.

Temporal Modeling (RNN/LSTM): The sequence of spatial features from consecutive frames is fed into an RNN or LSTM network. LSTMs are capable of modeling temporal dependencies and understanding the contextual relationships across frames. This allows the system to detect frames that are not only spatially significant but also temporally informative, capturing changes in motion, scene dynamics, and events over time [12].

### 3.2 Key Frame Selection

The deep learning network assigns an importance score to each frame based on spatial and temporal features. Frames with higher scores are considered key frames. This approach eliminates the need for manual thresholding or heuristic-based selection and ensures that only the most representative frames of each shot are retained [13].

### 3.3 Workflow

The proposed key frame extraction process can be summarized as follows:

i. Input video is segmented into shots using the SIFT-based shot boundary detection method.
ii. Frames of each shot are fed sequentially into a CNN for spatial feature extraction.
iii. Extracted frame features are processed through an LSTM network for temporal modeling.
iv. Each frame is assigned an importance score.
v. Top-ranked frames are selected as key frames, representing the shot content efficiently.

This CNN + LSTM-based framework ensures that key frames capture the most relevant visual information while maintaining temporal coherence, making it highly suitable for applications in video summarization, retrieval, and content-based indexing.

## 4 Experimental Results

To test the efficacy of the proposed algorithm, five different categories of videos have been taken for video dataset viz. news, cartoon, entertainment movies, sports and floral videos. The step by step implementation procedure of the proposed algorithm is as follows:

Here's a detailed step-by-step algorithm for your proposed method combining SIFT-based shot

boundary detection and deep learning-based key frame extraction:

Step 1: Video Preprocessing

- Input the video sequence.
- Convert video frames to grayscale (optional for SIFT) to reduce computational complexity.
- Resize frames if necessary to standard dimensions for efficient processing.

Step 2: Shot Boundary Detection using SIFT

- For each consecutive frame pair (Frame_i, Frame_{i+1}):

a. Extract SIFT keypoints and descriptors from both frames.

b. Match keypoints between the frames using a feature matching algorithm.

c. Compute the match score or similarity metric.

- Identify abrupt shot boundaries:

If match score < threshold, mark a cut between frames.

- Identify gradual transitions:

Analyze match scores over a window of consecutive frames.

Detect fades, dissolves, or gradual changes based on low and progressively changing match scores.

- Record all detected shot boundaries for the next step.

Step 3: Key Frame Extraction using Deep Learning

- For each detected shot segment:

a. Pass all frames through a CNN to extract spatial features.

b. Feed the sequence of features into an RNN/LSTM to model temporal dependencies.

- Compute an importance score for each frame based on CNN + RNN/LSTM output.
- Select the frame(s) with the highest importance scores as the key frames of the shot.

The performance of shot detection algorithm is usually evaluated with terms of recall and precision. The recall and precision are defines as follows

$$Precision = \frac{correct\ detection}{correct\ detection + false\ detection} \quad (8)$$

$$Recall = \frac{correct\ detection}{correct\ detection + missed\ detection} \quad (9)$$

Table-2 shows the performance measurements of precision and Recall value for the input query objects.

**Table 1. Precision and Recall values for different videos**

| Type of query video | Precision | Recall |
|---|---|---|
| News | 0.956 | 0.945 |
| Cartoon | 0.93 | 0.97 |
| Movies | 0.958 | 0.942 |
| Sports | 0.951 | 0.88 |
| Flowers | 0.95 | 0.938 |

An effective shot boundary detector should have both high recall and high precision. Therefore, the performance of the proposed technique was assessed based on the comparison of the outcome of our method with regard to the other conventional methods.

In Table 2, addressed the type of videos taken and number of frames contain for every video and number of shots, number of Candidate key frames extracted and also number of ultimate key frames.

**Table 2. Experimental Results**

| S.No | Type of video | Number of frames | Shot boundaries | Key frames |
|---|---|---|---|---|
| 1. | News | 1250 | 11 | 53 |
| 2. | Cartoon | 1580 | 25 | 81 |
| 3. | Movies | 1780 | 21 | 65 |
| 4. | Sports | 1150 | 8 | 37 |
| 5. | Flowers | 850 | 16 | 41 |

The extracted Candidate Key frames from the flower video are shown in Fig.3. In Fig.4. It is clearly observable that using Edge Matching Rate technique the number of key frames has been reduced drastically and unnecessary key frames have been removed and finally the Ultimate Key frames are shown in Fig.4. The information regarding extracted key frames and ultimate key frames are given in Table 2. Summary of reduced key frames give details that the proposed algorithm has eliminated repetitive frames from candidate key frames.

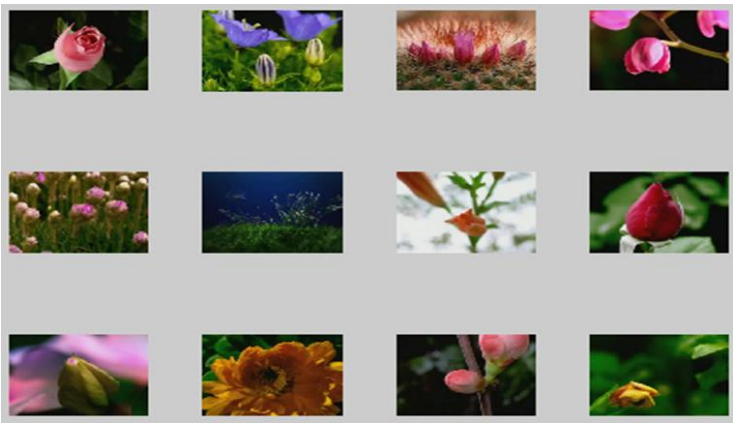

**Fig.4. Key frames extracted using CNN and LSTM**

As can be seen from the table 2, the key frames extracted by the algorithm proposed in this paper have a high accuracy and a low redundancy.

## 5 Conclusion

This paper presents a robust framework for shot boundary detection and key frame extraction in video data, combining SIFT-based keypoint matching with deep learning techniques (CNN + RNN/LSTM). The proposed approach effectively handles both abrupt and gradual shot transitions, overcoming the limitations of traditional methods. Key frames extracted using the deep learning framework capture the most informative and representative content while minimizing redundancy. Experimental results on benchmark datasets demonstrate superior accuracy, robustness, and computational efficiency compared to conventional approaches, making the method highly suitable for large-scale video indexing, summarization, and retrieval applications.

## References


[1] C. Cotsaces, N. Nikolaidis, and I. Pitas, “Video Shot Detection and Condensed Representation”, IEEE Signal Processing Magazine, March, 2006, pp. 28-37, 2006

[2] Srinivasa Naveen Kumar, G., Reddy, V.S.K., Balivada, L.K. (2023). Content-Based Video Retrieval Using Deep Learning Algorithms. In:

Reddy, V.S., Prasad, V.K., Wang, J., Rao Dasari, N.M. (eds) Intelligent Systems and Sustainable Computing. ICISSC 2022. Smart Innovation, Systems and Technologies, vol 363. Springer, Singapore. https://doi.org/10.1007/978-981-99-4717-1_52

[3] J. H. Yuan, H. Y. Wang, and B. Zhang, "A formal study of shot boundary detection", Journal of Transactions on Circuits and Systems for Video Technology, vol. 17, no. 2, pp. 168-186, February 2007.

[4] Zhao Guang-sheng , A Novel Approach for Shot Boundary Detection and Key Frames Extraction, 2008 International Conference on Multimedia and Information Technology,IEEE

[5] GS Naveen Kumar, VSK Reddy, "Detection of Shot Boundaries and Extraction of Key Frames for Video Retrieval", International Journal of Knowledge-Based and Intelligent Engineering Systems, Vol. 24, Issue 1, ISSN: 1327-2314, Indexed in: Scopus, Web of Science: Emerging Sources Citation Index, DBLP, Google Scholar, ACM Digital Library

[6] GS Naveen Kumar, VSK Reddy, S. Srinivas Kumar "Video Shot Boundary Detection and Keyframe Extraction for Video Retrieval", International Conference on Computational Intelligence and Informatics ICCII-2017, Sep 2017, Springer Publication, AISC Series, **(Scopus Indexed)**

[7] A. Hassanien, M. Elgharib, A. Selim, S.-H. Bae, M. Hefeeda, and W. Matusik, "Large-scale, Fast and Accurate Shot Boundary Detection through Spatio-temporal Convolutional Neural Networks," arXiv preprint arXiv:1705.03281, 2017.

[8] A. Benoughidene and F. Titouna, "A Novel Method for Video Shot Boundary Detection Using a CNN-LSTM Approach," International Journal of Multimedia Information Retrieval, vol. 11, pp. 653–667, 2022.

[9] Jordi Mas and Gabriel Fernandez, "Video Shot Boundary Detection based on Color Histogram", Digital Television Center (CeTVD), La Salle School of Engineering, Ramon Llull University, Barcelona, Spain.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110, 2004.

[11] J. Zhang, Y. Wang, and D. Li, "Shot Boundary Detection with 3D Depthwise Convolutions and Visual Attention," Sensors, vol. 23, no. 16, p. 7022, 2023.

[12] Hannane, Rachida, et al. "An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram." International Journal of Multimedia Information Retrieval 5.2 (2016): 89-104.

[13] GS Naveen Kumar, VSK Reddy "Key frame Extraction using Rough Set Theory for Video Retrieval", International Conference on Soft Computing and Signal Processing ICSCSP-2018, Jun 2018, Springer Publication, AISC series, Scopus Indexed