International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

An Effective Scheme for Shot Boundary Detection and Key Frame Extraction for Video Retrieval

Gs Naveen Kumar, Vsk Reddy

Abstract**: *The rapid development of devices for image capture and information sharing has resulted in the availability of huge amounts of online video for various applications such as education, news, entertainment, etc. This leads to problems and difficulties when users query any content-related video. The reason for this scenario is that the presently available techniques of content representation and retrieval are based primarily on annotation. It therefore provides insufficient information for understanding and retrieving the content to match the query of the user. Content Based Video Retrieval (CBVR) is one of the promising new ways for finding content in a large video archive, rather than simply searching terms. The primary steps for indexing, summarizing and retrieving video are shot transition recognition and representative frame extraction. We have proposed a key point matching algorithm for a superior and robust Scale Invariant Feature Transform (SIFT) followed by the collection of representative frames from each segmented shot using the Image Information Entropy method. By using the Rough Set Theory, we can get better the concert of this scheme through removing unnecessary representative frames. All the methods suggested to prove the efficacy were tested on TRECVID datasets and contrasted with state-of - the-art approaches.***

KEY WORDS: Shot Transition, Representative frames, Image Entropy, SIFT, Rough Set Theory, Retrieval.

I. INTRODUCTION

Video content-based access to information is a method that uses aesthetic components to browse images from sources of large-scale image data in terms of interest rates for individuals. Shade, shape, structure, as well as spatial format for indexing and identifying the image are the aesthetic materials of a picture used in Aesthetic Content-based Details Access. Aesthetic Content-based Info access is utilized in certain domain names to find matching in instance of unlawful photo usage as well as determining bad guys from iris as well as finger prints picture.

There is a huge growth of digital data modification annually. Every year, huge amounts of sound and visual information are produced by electronic camera surveillance, TELEVISION programs and house camera as well. World Wide Web (WWW) production makes this electronic information available worldwide. A high amount of audio-visual information makes it practically impossible to surf data. Numerous storage space solutions are available, such as Compact Disk Review Only Memory (CD-ROM) and Digital Versatile Disk (DVD), but the level of availability they provide is much lower.

It must always be ensured that the various methods of arranging video must remain in sync with the tremendous amount of production of data. There is therefore an urgent need for much better techniques of entry. In reality, the inefficiency and weakness of traditional approaches used for VR has led to the need for brand-new strategies that can change the content-based video data source. The CBVR is therefore considered a demanding task, which is a multidisciplinary knowledge access server project. Every day, the consumer demand for esthetic data is growing.

As a result, advanced innovation is needed to support, model, index and also recover multimedia information. There is a need for comprehensive approaches for accessing visual details. Content-based video recovery is careful to be an unpredictable mission. The fundamental intention at the back of this is the measure of intraclass divergence where the indistinguishable semantic idea happens under different conditions like light, appearance, and scene settings. For example, recordings involving a man riding a bike can have inconsistency as different sizes, appearances, and camera movements. The greater part of the exploration in the zone of substance based video recovery is implied at manage these difficulties. Therefore, different viewpoints required to be meticulous to settle on whether two videos are practically identical or not while investigating the video content. Besides, understanding video substance is ordinarily a skewed strategy for a customer. Marking video data with a predefined set of names altogether smoothens the advance of pursuit. It is not anticipated that it would catch all encouraging perspective purposes of clients. The resulting identification of specific video material in the advanced video's exponentially accumulating test is passed for an intense errand. An allegorical form is exposed in Hun-Woo Yoo et al. (2006) throughout the liquifying cycle. The outcomes of the LEB are motivating, it is delicate to cam movement, item movement as well as a substantial material modification within the shot [2] as specified by Hun-Woo Yoo et al. (2006).

Manuscript published on November 30, 2019.
* Correspondence Author

Gs Naveen Kumar*, Research Scholar, Faculty of Engineering, Lincoln University College, Malaysia

Vsk Reddy, Professor, Faculty of Engineering, Lincoln University College, Malaysia.

These issues can be managed by utilizing the side attributes, considering that side attributes are durable to lighting, cam activity, things activity and also substantial material modification within the shot. The revolutionary approach is proposed to improve performance and also to overcome the problems that exist in the LEB. The downside is that if two different frames have the same color range, it's a mistake. Several key frame [3] identification techniques to increase the speed of Content Based Video Retrieval Systems are found in the literature and some of the relevant techniques are presented here. Che-Yen et al. (2007) extracted features such as motion and color information and stored them as a feature vector and selected the key frames by examining the temporal variation of the feature vector trajectory.

In this paper, we proposed a pioneering technique for scanning boundary detection using on Scale Invariant Feature Transform (SIFT) [4] to defeat the shortcomings of parallel research algorithms. It is stable and invariant regardless of the rotary motion of the picture, the scale of the image, the noise and the lighting variation. Main frame extraction helps captured boundary detection in an object description. The extraction of the main frame is selective or descriptive extraction of the frame. This removes unnecessary information and decreases the amount of data needed to search and retrieve images. The entropy values for all frames in the shot are determined in the proposed algorithm. Frames that have dissimilar entropy rates are considered to be representative frames. There is a probability of similar representative frames being replicated in dissimilar shots [12] [13]. These repetitive representative frames are called redundant representative frames, which direct to enlarge in video illustration information. Eventually, the Rough Set Theory [5] is used to remove redundant key frames and to arrive at the final key frame. The description of the left over part is arranged as follows[19] [20]: Shot Boundary Transitions is talk about in Section 2, Key Frame Extraction is talk about in Section 3, Structure of the proposed method is discussed in Section 4, the experimental outcomes are describe in Section 5 and the conclusion is illustrate in Section 6.

II. SHOT BOUNDARY DETECTION USING PROPOSED SIFT FEATURE

This section presents an overview of the video components and discusses the generic framework available for automatically detecting shots in the video. Based on a review of a number of literature related to Shot Boundary Detection (SBD) [17], a new approach to SBD has been proposed to ensure fair trade between accuracy and time of computation. Eventually, studies have been carried out to improve the efficacy of the proposed methods in the scheme. Scale Invariant Feature Transform (SIFT) [7] was designed in this paper to extract video features.

2.1 SCALE_INVARIANT FEATURE TRANSFORM:

Lowe's approach (Lowe 2004) for photo accent era converts a photo right into a large build-up of accent vectors, each of which is stable for photo analysis, scaling as well as pivoting, primarily standard for information modifications and also passionate for adjacent geometric contortion. These highlights share the relative residential or commercial properties of nerve cells in crucial aesthetic cortex, which enshrine essential frameworks, shielding and also growth in the field of primate vision.

Hidden positions are known as the maximum as well as the minimal impact of the difference in Gaussian capacity added to the creation of smoothed as well as remodeled images in the range space. Filter descriptors Hearty to the neighboring family member contortion are after that obtained by believing about pixels moving the critical spot, covering and also re-sampling the group photo instructions aircraft.

The following are the major steps in the calculation used to produce the set of image attributes:

1. *Scale-Space extrema discovery:* This phase searches various ranges and also picture places utilized for extrema discovery. Gaussian distinction feature is made use of to determine hidden factors of problems that are fit to range as well as instructions.

$$D(m,n,\sigma) = \left(G(m,n,k\sigma) - G(m,n,\sigma)\right) * I(m,n)$$

(2.1)

$$= L(m,n,k\sigma) - L(m,n,\sigma)$$

$$L(m,n,\sigma) = G(m,n,\sigma) * I(m,n) \qquad (2.2)$$

$$G(m,n,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(m^2+n^2)/2\sigma^2} \qquad (2.3)$$

2. *Perfect Key Point Localization*: At each prospect area, an in-depth version is fit to establish place as well as range. Bottom line are chosen based upon steps of their security.

3. *Orientation Assignment*: Each area based bottom line on neighborhood picture slope instructions will certainly be set aside with even more instructions. The picture information that has actually been changed loved one to the designated instructions, range, as well as place for each function.

$$M(m,n) = \sqrt{\left(L(m+1,n) - L(m-1,n)\right)^2 + \left(L(m,n+1) - L(m,n-1)\right)^2}$$

(2.4)

$$\theta(m,n) = \tan^{-1}\left(\frac{L(m,n+1) - L(m,n-1)}{L(m+1,n) - L(m-1,n)}\right) \qquad (2.5)$$

4. Descriptor Representation

A Descriptor for each key point is produced commencing local image gradient information. This procedure forms a vector descriptor for SIFT [8] features viewing 128 directions in Figure 1.

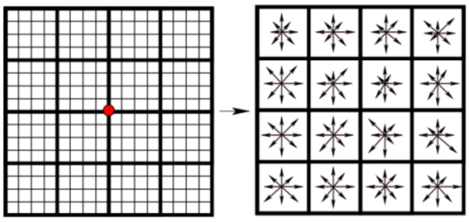

Figure 1: Key point descriptor creation using128 directions

2.2 Key Point Corresponding for Shot Transition Detection

FILTER keypoints are taken from the structures of the video clip as well as after that the ratios of the matched keypoints number to the total number between the structure iand also the structure i+N are used to detect the limits of the shot. Number 2 shows the corresponding local role between two frameworks.

Retrieval Number: D5436118419/2019©BEIESP
DOI:10.35940/ijrte.D5436.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The crucial factors in the present structure are contrasted with those of the following framework by making use of the k-Nearest neighbor (kNN) Look (Aburomman et al. 2016) [1].

$$d(a,b) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2} \qquad (2.6)$$

When the crucial factors in two structures are matched with each various other, lines are attracted in between the matching vital factors. If the number of matched essential factors in between any kind of two frameworks is better than the limit worth, these 2 structures are relevant to the very same area.

III. KEY FRAME EXTRACTION

It is a significant video recovery process. Also a precise or representative frame is called a key frame. This includes the shot's required data and review the whole content of the remaining frames. This removes unnecessary information and increases data amounts for indexing and video retrieval.

3.1 Image Information Entropy

For removing key frameworks, information entropy of each framework in the shot is determined [8] [9]. The details decline calculation for a structure is revealed in Formula (3.1)

$$Entropy = -\sum_{i=1}^{L} p(x_i) \times \log(p(x_i)) \qquad (3.1)$$

Where,

$$p(x_i) = \frac{total\ (x_i)}{m \times n} \qquad ,$$

$$0 \leq p(x_i) \leq 1 \ and \ \sum_{i=1}^{n} p(x_i) = 1$$

The block diagram of key frame extraction is shown in figure 2.

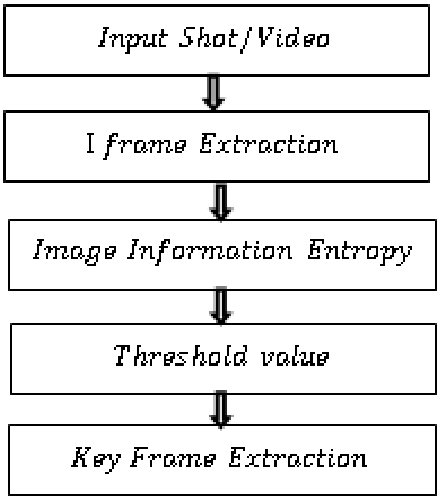

Figure 2: Block Diagram of Key Frame Extraction

3.2 Extraction of Ultimate Representative frames using and Rough Set Theory

It is observed that sometimes the object and also history repeat in various shots of the video, e.g. an information viewer narrating news story, replay in sporting activities, lecture videos, etc.. This results in several redundant representative frames[6] [10]. To get rid of these repetitive crucial frameworks, a filtering system action is carried out, where each representative frame is compared to every other representative frame to discover the duplicate or near-duplicate structures[14][15].

The implementation of this algorithm is shown in the figure 3.

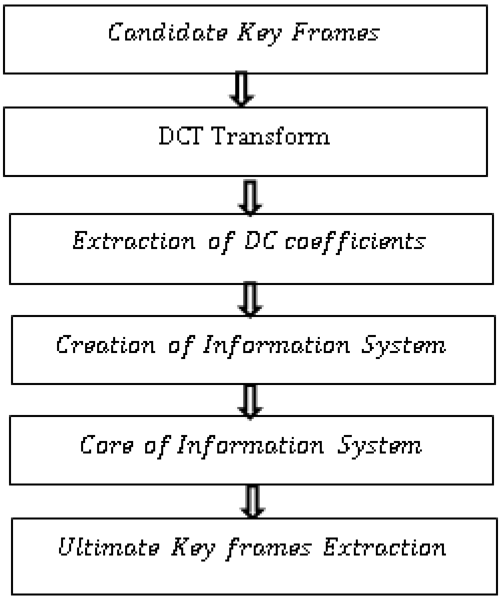
Candidate Key Frames
DCT Transform
Extraction of DC coefficients
Creation of Information System
Core of Information System
Ultimate Key frames Extraction

Figure 3: The framework for Ultimate Key frame Extraction using RST

3.3 Rough Set Theory (RST)

Roughset theory [11] is a powerful mathematical tool to rectify imperfect knowledge. Now-a-day, many researchers face the problem of imperfect knowledge in the areas of computer applications, robotics, machine learning, artificial intelligence, signal processing by manipulating imperfect knowledge into perfect knowledge. The procedure for extracting the Core set is as follows [16] [18].

1. Let G be a subset of AR. Equivalence relation of set G needs to be found to arrive at Indiscernibility of G. It is denoted by IND(G).

2. If IND(G)=IND(G-{a}), a∈G then **a** is dispensable i.e**a** can be done away with. If IND(G)≠IND(G-{a}), then a is indispensable i.e. **a** is an essential attribute.

3. Let M represent a proper subset of G i.e. M⊂G. We need to find Indiscernibility of M i.e. IND(M).

4. If IND(M)=IND(G) even after subtracting an attribute or a few attributes from G, then M is called a reduct of G.

5. We find all the reducts in G with different combinations of attributes M. The set of all reducts in set G is denoted by RED(G).

6. The set of all indispensable attributes in G is called the Core of G and it is denoted by CORE(G), where Core of G is the intersection of Reducts in G; CORE(G)= ∩ RED(G).

The core set contains major visual color and motion information and does away with the redundant video information.

IV. EXPERIMENTAL RESULTS

The step by step process to execute the projected scheme as shown in Figure 4

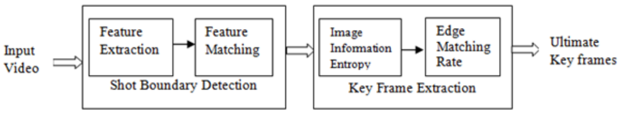

Figure 4: Framework for the proposed method

In order to check the efficiency of the algorithms: Shot Transition discovery using SIFT, Selective Frame Extraction using Frame data Entropy and Rough Set Theory, five genre video categories are taken for video dataset viz. Entertainment, Sports, Cartoon, News, Commercial. Shot transition recognition output using SIFT algorithm is evaluated for Precision and Recall is defined as follows.

$$Precision = \frac{correct\ detection}{correct\ detection + false\ detection} \times 100 \qquad (4.1)$$

$$Recall = \frac{correct\ detection}{correct\ detection + missed\ detection} \times 100 \qquad (4.2)$$

The experimental outcomes demonstrate that the approach suggested exceeds the existing methods as shown in Table-1.

Table 1. Accuracy and recall values of various video classes.

Type of query video	Block Matching Method		Histogram Based Method		Proposed Method	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Entertainment	79.7	78.3	84.3	82.4	95.7	94.4
Sports	77.9	74.1	80.7	79.1	94.4	92.2
Cartoon	76.6	73.8	82.6	80.2	94.9	93.2
News	78.3	75.7	81.5	78.7	95.2	91.5
Commercial	79.3	76.2	83.8	81.3	95.5	93.9

The type of videos taken, the number of frames in each video, the number of shots identified, the number of candidate key frames extracted and the number of ultimate key frames are shown in Table 2.

Figure 6: Main frames of the candidate extracted from the flower clip

Figure 7: Extracted the ultimate main frames using Edge Matching Rate

Table 2. The Extracted Shot Boundaries, Key frames and Ultimate Key frames.

S.No	The type of video	Number of frames	Shot boundaries	Candidate Key frames	Ultimate Key frames
1.	News	1240	12	59	24
2.	Cartoon	1480	26	71	51
3.	Movies	1790	22	93	53
4.	Sports	1140	9	42	29
5.	Flowers	880	13	39	31

Figure 5 displays the candidate main frames extracted from the video clip. Figure 6 shows the number of redundant key frames extracted by using the Rough Set Theory resulting in the final key frames.

V. CONCLUSION

A novel scheme has been projected in this chapter to detect shot boundaries using spatio-temporal Scale Invariant Feature Transform (ST-SIFT). The methodology is effective in shooting adjustments to lighting and camera / object movement. Six different video sources experiments show that the proposed ST-SIFT algorithm has yielded better detection quality than LEB. The drawback in the other algorithms is the extraction of unneeded representative frames. Redundant key frames have been reduced by use an Rough Set Theory in the projected algorithm. Accurate Shot transitions and Representative frames have been extracted by making utilize of these method so, this algorithm has revealed elevated precision and elevated recall rate. Hence, it is robust and effectual.

REFERENCES

1. Aburomman, A. A., & Reaz, M. B. I. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. Applied Soft Computing, 38, 360-372.
2. C. Cotsaces, N. Nikolaidis, and I. Pitas, “Video Shot Detection and Condensed Representation”, IEEE Signal Processing Magazine, March, 2006, pp. 28-37, 2006.
3. Chen, Ling, and Yuhong Wang. "Automatic key frame extraction in continuous videos from construction monitoring by using colour, texture, and gradient features." Automation in Construction (2017), Elsevier, 355-368.
4. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110, 2004.
5. GS Naveen Kumar, and V. S. K. Reddy. "Key Frame Extraction Using Rough Set Theory for Video Retrieval." In Soft Computing and Signal Processing, pp. 751-757. Springer, Singapore, 2019.
6. Hua, Guogang, and Chang Wen Chen. "Distributed video coding with zero motion skip and efficient DCT coefficient encoding." In Multimedia and Expo, 2008 IEEE International Conference on, pp. 777-780. IEEE, 2008.
7. Li, Jun, et al. "A divide-and-rule scheme for shot boundary detection based on SIFT." JDCTA 4.3 (2010): 202-214.
8. Lina Sun and Yihua Zhou, "A key frame extraction method based on mutual information and image entropy," 2011 International Conference on Multimedia Technology, Hangzhou, 2011, pp. 35-38Liu, Gentao, et al. "Shot boundary detection and keyframe extraction based on scale invariant feature transform." Computer and Information Science, 2009. ICIS 2009. Eighth IEEE/ACIS International Conference on. IEEE, 2009.
9. Luo, Y & Junsong Yua, 2013, Salient Object Detection in Videos by Optimal Spatio-Temporal Path Discovery. In proceedings of ACM International conference on Multimedia, pp. 509-512.
10. Nasreen, Azra, and G. Shobha. "Reducing redundancy in videos using reference frame and clustering technique of key frame extraction." In International Conference on Circuits, Communication, Control and Computing, pp. 348-440. IEEE, 2014.
11. Pawlak, Zdzislaw. "Rough set theory and its applications to data analysis." Cybernetics & Systems 29.7 (1998): 661-688.
12. Qu, Zhong, Lidan Lin, TengfeiGao, and Yongkun Wang. "An improved keyframe extraction method based on HSV colour space." JSW 8, no. 7 (2013): 1751-1758.Ren, Liping, et al. "Key frame extraction based on information entropy and edge matching rate." Future Computer and Communication (ICFCC), 2010 2nd International Conference on. Vol. 3. IEEE, 2010.
13. Saravanan, D., & Vengatesh, K. J. (2015). Video content reterival using historgram clustering technique. Procedia Computer Science, 50, 560-565.

14. Shirahama, Kimiaki, Yuta Matsuoka, and Kuniaki Uehara. "Event retrieval in video archives using rough set theory and partially supervised learning." Multimedia Tools and Applications 57, no. 1 (2012): 145-173.
15. Thakre, K. S., A. M. Rajurkar, and R. R. Manthalkar. "Video partitioning and secured keyframe extraction of MPEG video." Procedia Computer Science 78 (2016), Elsevier, 790-798.
16. Uehara, Takeyuki, Reihaneh Safavi-Naini, and Philip Ogunbona. "Recovering DC coefficients in block-based DCT." IEEE Transactions on Image Processing 15.11 (2006): 3592-3596.
17. Wu, Zhonglan, and Pin Xu. "Shot boundary detection in video retrieval." Electronics Information and Emergency Communication (ICEIEC), 2013 IEEE 4th International Conference on. IEEE, 2013.
18. Xu, Junyu, Yuting Su, and Qingzhong Liu. "Detection of double MPEG-2 compression based on distributions of DCT coefficients." International Journal of Pattern Recognition and Artificial Intelligence 27, no. 01 (2013): 1354001.
19. Yoo, HW, Ryoo, HJ & Jang, DS 2006, "Gradual shot boundary detection using localized edge blocks", Multimedia tools applications, vol. 28, pp. 283-300.
20. Zhao Guang-sheng , A Novel Approach for Shot Boundary Detection and Key Frames Extraction, 2008 International Conference on Multimedia and Information Technology,IEEE.