Volume 7, Issue 2, February 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Segmentation of Tri-Lingual Documents

[1]Mahesha D M, [2]Bhavya D N, [3]Nandini H M

[1]Department of Studies in Computer Science, Karnataka State Open University, Mysore,India

[2]Department of Studies in Computer Science, Karnataka State Open University, Mysore,India

[3]Department of Studies in information technology, Karnataka State Open University, Mysore, India

Abstract*— Physical layout analysis intends to study the arrangement of layouts or locations of the regions present in a document image before understanding it. Before extracting the text or information from a document image, page segmentation (layout analysis) techniques need to be applied to identify the exact layout (area) where the text or image resides. In Page Segmentation, Top-down methods are simple and efficient but fail in non Manhattan layouts. In contrast, Bottom-up approaches adapt non Manhattan layouts easily than the top down approaches, but heavily depend on the threshold, parameters and extensive computations for layout identification. On the other hand, Hybrid methods (Bruel [31], Bruel [32]) suits well for layout identification by eliminating the dependency on threshold and parameters. But this analyzes the white background of the image with small white rectangles and merges them to locate the content blocks. Merging of small white rectangles makes the identification process tedious since large number of small white rectangles gets involved in the image. In addition, this approach heavily relies on heuristics for merging operations, which affects the segmentation rate considerably. In all the above reported methods (Bottom up and Hybrid approaches), connected component analysis (requires more number of pixel visits) is required to identify black and white components from the image. Therefore, the above shortcomings motivated this research towards designing a White Space analysis technique which eliminates the usage of the connected component analysis (to identify white spaces), heuristics, threshold and prior knowledge. As a result, in this thesis, Rectangular White Space Analysis (RWSA) technique has been proposed to grab all the white spaces over the image in a single scan over the image with minimum pixel visits, and the white spaces are merged together without the assumptions of heuristics and threshold to segment the layouts. Moreover, two statistical properties have also been proposed in this thesis, to separate the text blocks and images from the identified layouts and this hybrid approach has been explained in the subsequent section.

***Keywords*— Segmentation; Section finding; Section Merge; Feature Extraction; Indexing**

I.INTRODUCTION

Anything which conveys information is known as a document. Generally, a document is a knowledge container. Most of the times we acquire knowledge from documents such as Newspapers, Textbooks, Scientific journals, Magazines, Technical reports, Office files, Postal letters, Bank cheques, Application forms etc. (Tang et al., [1]). To understand the huge information, an extensive amount of manual processing is required and such a manual processing is very much time consuming. To overcome this difficulty, it is essential to automate the manual process which needs efficient algorithms. This automation process is considered as document image processing (DIP). In general, the document image processing is divided into text processing and graphics processing. Text processing is further divided into character recognition and page layout analysis. Graphics processing is further divided into line processing and region processing as shown in Figure 1.

A. Stages in Document Image Processing

The document image processing involves three basic steps at conceptual levels, which are document image analysis, document image recognition and document image understanding. Within these three levels, there are several other interacting modules such as image acquisition, binarization, block segmentation, block classification, logical block grouping, character and word recognition, picture processing and analysis, graphic analysis, picture understanding, text understanding and graphics understanding. The interactions between these processes and data flow between levels are shown in Figure 2.

1) Document Image Analysis

Document image analysis is a process of recovering syntactic and semantic information from images of documents, prominently from scanned versions of paper documents. There are two distinct tasks in document image analysis. The first has a syntactical goal consisting of the identification of basic components of the document, the so-called document objects. The second has a semantic goal consisting of the identification of the role and meaning of the document objects in order to have an interpretation of the whole original document. The structural analysis, on the other hand involves usage of layout clues to identify headlines, locate different lines, etc. In general, image analysis involves

the extraction and use of attributes and structure relationships in the document in order to label its components within contextual rules dictated by the document class. Analysis of printed documents obviously involves skew angle estimation and correction which is a very challenging task.

2) Document Image Recognition

Document Image Recognition (DIR), a very useful technique in office automation and digital library applications, is to find the most similar template for any input document image in a prestored template document image. Nowadays a large amount of existing paper documents are transformed to digital document images through scanners and cameras. However, the next step is to analyze a document and segregate text blocks, graphic block, picture block, etc, so as to facilitate labeling of the blocks. This process of labeling the blocks is said to be document image recognition or identification.

3) Document Image Understanding

Document image understanding is a component which extracts the logical relationships between the respective blocks of a document. Logical document structure is a hierarchical representation of semantics of the given document. The same logical document structure is formatted in varieties of physical layouts by changing the variables such as number of pages and font sizes, spacing between paragraphs and between sections, number of columns, etc. In all these layouts, the semantics of the document remains unaltered. Logical structure analysis determines the document‘s semantic structure and provides data appropriate for information retrieval.

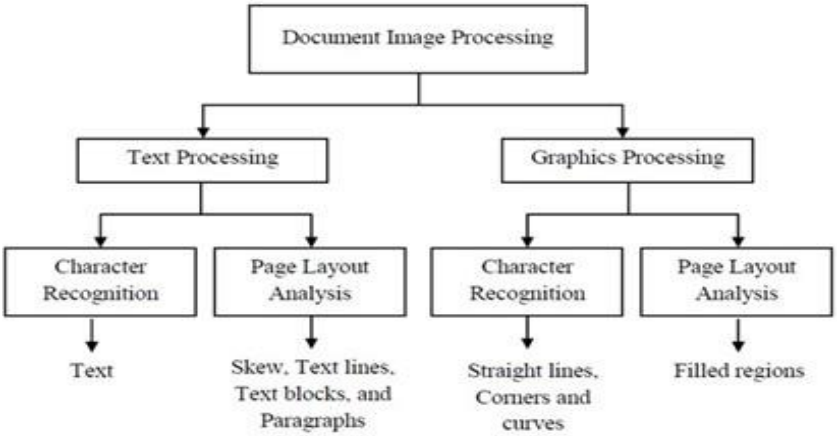


Figure 1. Hierarchy of document image processing with subcategories

B. Script Recognition

The OCR technology for Indian documents is in emerging stage and most of these Indian OCR systems can read the documents written in only a single script. As per the Indian constitution, every state Government has to produce an official document containing a national language (Hindi), official language (English) and state language (or regional language). According to the three-language policy adopted by most of the Indian states, the documents produced in an Indian state Karnataka, are composed of texts in the regional language- Kannada, national language Hindi and the world wide commonly used language-English. In addition, majority of the documents found in most of the private and Government sectors of Indian states, are Trilingual type (a document having text in three languages). So, there is a growing demand to automatically process these Trilingual documents in every state in India, including Karnataka

The monolingual OCR systems will not process such multi-script documents without human involvement for delineating different script zones of multi-lingual pages before activating the script specific OCR engine. The need for such manual involvement can result in greater expense and crucially delays the overall image to text conversion. Thus, an automatic forwarding is required for the incoming document images to handover this to the particular OCR engine depending on the knowledge of the intrinsic scripts. In view of this, identification of script and/ or language is one of the elementary tasks for multi-script document processing. A script recognizer, therefore, simplifies the task of OCR by enhancing the accuracy of recognition and reducing the computational complexity.

Script Recognition approaches can be broadly classified into two categories, namely, local and global approaches. The local approaches(Pal and Chaudhury [2], Pal et al [3]) analyze a list of connected components (Line, word, char) in the document images, to identify the script(or class of script). In contrast, global approaches (Joshi [4]) employ an analysis of regions (block of text) comprising atleast two lines (or words)without finer segmentation. In general, global approaches work well based on texture measurement, but this relies heavily on a uniform block of text (Buschet al [5]), and extensive preprocessing (to make the text block uniform) is required to measure the texture. Even though local approaches rely on the accuracy of character segmentation or connected component analysis, it could work well on the documents irrespective of their quality or uniformity in the block of text.

In the literature, many works have been reported for script recognition at the document, line and word levels, using local approaches. In this context, researchers have made a number of attempts to discriminate the Han and Latin script (Spitz [6], Lu and Tan [7]) at the document level and

exploited many Indian scripts at line level and word level (Pal and Chaudhury [8], Pal and Chaudhury [9], Padma and Nagabhushan [10], Dhandra et al [11]). However, all the techniques reported in the literature are script dependent. Since this research is intended to develop an classification system for kannada document images, Script Recognition, to discriminate the kannada from English scripts in bilingual document images is becoming important. In this connection, few local approaches are reported in the literature, such as spatial spread analysis (Dhanya et al [12]), Aspect Ratio (Tan et al [13]), Structural features (Pal and Chaudhury [14]), and Water Reservoirs (Pal et al [15]). However, all the above mentioned techniques produce a low discrimination rate due to its incapability in exploration of thescripts. Global approaches (Pati et al [17], Pati and Ramakrishnan [18]). S Chaudhury et al., [19] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu.

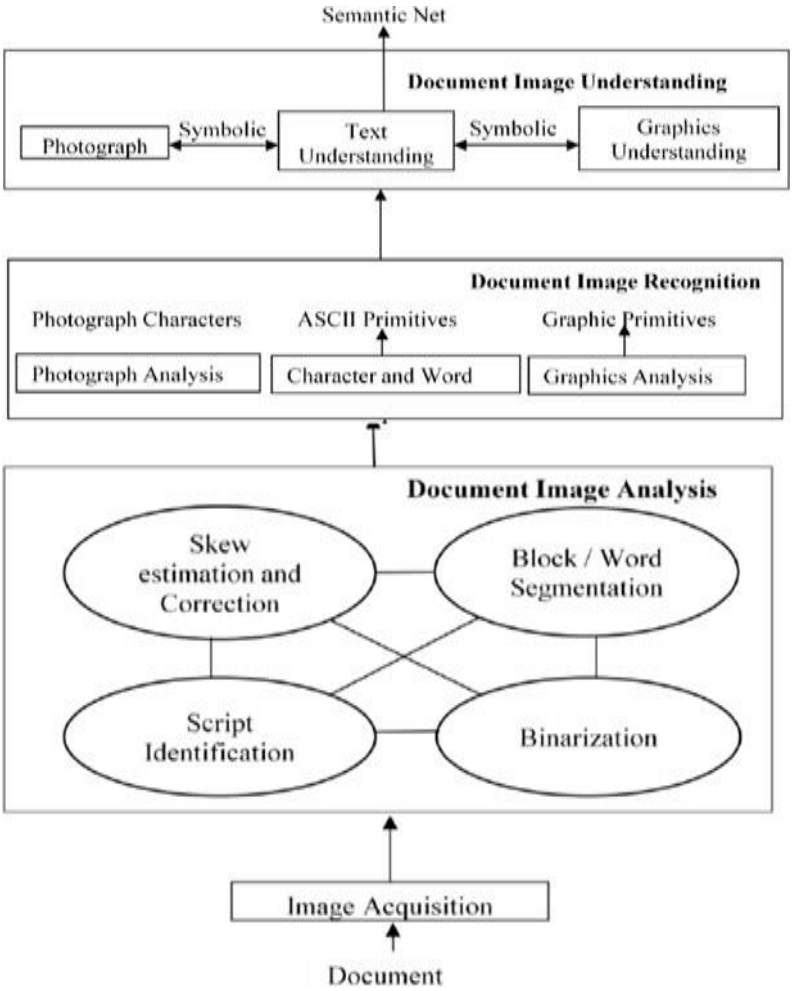


Figure 2. Steps involved in document image processing

G D Joshi et al., [20] have presented a script identification technique for 10 Indian scripts using a set of features extracted from logGabor filters. Dhanya et al., [21] have used Linear Support Vector Machine (LSVM), K-Nearest Neighbour (K-NN) and Neural Network (NN) classifiers on Gabor-based and zoning features to classify Tamil and English scripts. Hiremath [22] have proposed a novel approach for script identification of South Indian scripts using wavelet based co-occurrence histogram features. S R Kunte and S Samuel [23] have suggested a neural approach in on-line script recognition for Telugu language employing wavelet features. Peeta et al., [24] have presented a technique using Gabor filters for script identification of Indian bilingual documents

II. SEGMENTATION

A. System Method

A hybrid layout analysis framework along with a text image separation technique has been proposed in this thesis to identify simple and complex images and to separate the text and images from the layouts. The impact of the hybrid layout analysis has been motivated in this thesis to design a Rectangular White Space Analysis technique, since white spaces are a generic layout delimiter. In general, publishers use white spaces to separate blocks of text since they are constrained by common printing technology. The RWSA technique is parameter free and threshold free and adapts to various heterogeneous structures through layout gaps rather than connected component analysis and heuristics. Also, two Statistical properties such as the Black run length, and Transition rate have been proposed in this thesis to isolate the textual areas from images in the segmented blocks.

The architecture of the Hybrid layout analysis using RWSA is depicted in Figure 3.1. This kind of a layout analysis is necessary before understanding the text in all applications. When document images are given as input to this system, they undergo Noise removal and Binarization in the preprocessing phase.

RWSA technique consists of White space Section Finding, Section Merging, Cropping Extraneous section and Rectangular Formation phases to identify layouts. The input image to the system could be a color or a gray scale image obtained by scanning the newspapers. Documents containing text, graphics, figures, maps and tables are taken as input. Scanned input images passes through the following phases to identify the layouts.

1) Pre-processing

The preprocessing of the document images includes the process of Noise removal and Binarization. Noise could occur in the document images due to many sources such as aging, photocopying etc. and the application of filters reduces noises in these images. Here noise has been suppressed in the document image by using a median filter, since median filters smear the character image strokes. For this median filter, a 3*3 mask has been chosen and it is applied over the image, which replaces nine pixels by the intensity of the center pixel over this mask. As a result of pulling the median filter output to the gray level of the center pixel, the shapes of the character strokes can be preserved. Binarization has been applied after noise removal. Binarization is a technique by which the color and gray scale images are converted into binary images. The most common method is to select a proper threshold for the image and convert all the intensity values above the threshold into an intensity value representing as _white‘ and below the threshold as _black‘ value. All intensity values below a threshold are converted to one intensity level and intensities higher than this threshold are converted to the other chosen intensity.

2) Section Finding

After noise removal, RWSA technique has been applied over the image which contains Section Finding as the initial process. Since all the books and magazines use white spaces as a separator within and between the texts, the observation of small white spaces becomes mandatory to identify the text area. Therefore, in Section Finding, white spaces are used as delimiters and observed for analysis. Variable length white spaces exist inside the text in both the directions, apart from the white spaces surrounding the textual zones. Due to the existence of non-uniform, small white gaps in the image apart from the column separators, a careful analysis is required to observe and record the white spaces (There may be a possibility of a small white space separator inside and across the paragraphs). As a result, in this thesis, the width of the image has been divided into _n‘ equal sections (The Total number of sections is defined as the ratio of the Width of the image to the Section length; the Section length is experimentally fixed as 5 pixels and this length suits all kind of images). Since connected component analysis has been eliminated, a single horizontal scan has been performed over the image to grab the white spaces. After an entire horizontal scan of an image, all the sections which appear as white spaces are reported and their positions with the corresponding row number (section numbers along with their row numbers) have been recorded as a result of this procedure.

3) Section Merging

It is hard to process various white space section numbers to identify the layout gaps if the merging procedure has been avoided. Once all the white space section numbers based on their row number have been indicated, the merging of adjacent sections in both the directions is required to form horizontal and vertical white space rectangles which are done through the Section Merging phase.

The Section Merging phase consists of two processes: Horizontal Section Merging and Vertical Section Merging. Initially, horizontal section merging accepts all the white space section numbers with their corresponding row numbers as the input and produces a series of within-line or row-wise white space clusters as output(i.e.), subsequent white space sections in each row gets merged together to produce a series of row-wise white space sections. Since all the white spaces (section-wise) are identified and merged properly, the chance of getting under-segmentation has been completely eliminated.

B. Rectangular Analysis

The rectangular analysis phase consists of Cropping and the Rectangular formation process. After the identification of horizontal and vertical white space rectangles, finding the areas which are uncovered by the white space rectangles could yield the layout. Deviated edges or edges which do not have intersections over them, in horizontal and vertical white space rectangles must be trimmed or cropped properly to obtain the areas which are uncovered by the white spaces. In order to crop the non intersecting portion of the edges, a cropping procedure has been designed and applied over the horizontal and vertical sides of the rectangles.

1) Cropping Procedure

The Cropping procedure acts over the white space rectangles in both the directions by accepting the horizontal edges of each Horizontal White Space Rectangle (HWSR) and the vertical edge of each Vertical White Space Rectangle (VWSR). In a horizontal orientation, this procedure attempts to identify the two boundary vertical edges which pass through the edges of each HWSR and crops the extraneous portion of the horizontal edges of each HWSR, which appears apart from the intersecting boundary of the vertical edges. If no two boundary vertical edges pass through the horizontal edge of the HWSR, then the total horizontal edge would be removed for further processing.

2) Rectangular Formation

Once the horizontal and vertical edges are cropped, the areas uncovered by the white spaces could be easily extracted through rectangular formation procedure. Recursively, this procedure takes two cropped horizontal edges of each HWSR and checks with each pair of the vertically cropped edges for the formation of a rectangle. If the integration of these pairs of horizontal and vertical edges coincides,

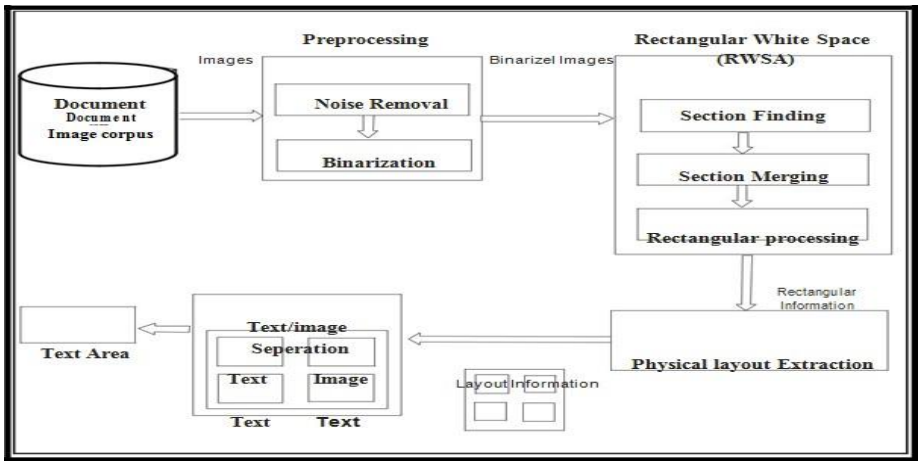


Figure 3 Hybrid Layout Analysis Architecture

Figure 3. Input Image with text

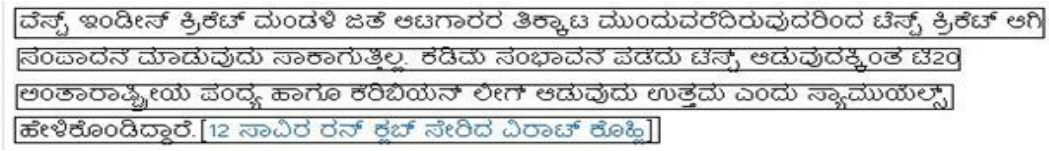


Figure 4. Input Image with elimination text

C. Text/ Image Analyzer

Once the content blocks have been identified, the next step attempts to separate the textual blocks from the images and pictures, since textual blocks are required for further processing. Once the homogeneous regions are obtained, each region gets passed into the text image analyzer to identify the text component.

Two statistical properties called as Black Run Length (BRL) and White Black Transition Count (WBTC), which spans in the horizontal direction of the image have been used here to identify the textual blocks. Black run length corresponds to the ratio of the total number of black pixels in a row to the total transition (black-white disposition) count in that row. The White Black Transition count corresponds to the ratio of the total number of transitions in a row to the total number of pixels in that row. It is concluded that if the mean black run length appears to be more, and the Mean white black transition count of all the rows appears to be lesser than the threshold, it is concluded as image and not as a text. The Black Run and Transition count of each row in the region is computed as stated in equations (3.1) and (3.2).

The Black run BR (black density), of the ith row of the region is represented in equation (3.1) and Transition count TC of the ith row of the region has been computed as stated in equation (3.2)

$$BR(f(x,y)) = \begin{cases} br = br+1 & \text{if } f(x_i, y)=0 \\ br & \text{otherwise} \end{cases} \quad (3.1)$$

$$TC(f(x_i, y)) = \begin{cases} tc\ tc\ 1 \text{ if } (f(x_i, y)\ 1)\&((f(x_{i\ 1}, y))\ 1 \| (f(x_{i\ 1}), y)\ 1) \\ Tc & \text{otherwise} \end{cases}$$

where x_i represents i^{th} row and $(x1\ x_i\ \ x2)$ (3.2)

where f(x,y) is the 2-D array of the pixel region with the coordinates ranging from $x_1,y_1 \ldots x_2,y_2$ and BR represents black run , TC represents the transition count, 1 represents the presence of a white pixel and 0 represents the presence of a black pixel.

The Black Run Length (BRL) of the ith row has been computed as the ratio of Black runs in a row to the total transition count in that row as stated in equation (3.3) and White to Black Transition Count (WBTC) of the ith row corresponds to the total number of transitions in that row to the total number of pixels in that row.

$$BRL_i(f(x_i, y)) = \frac{BR(f(x_i, y))}{TC(f(x_i, y))} \quad (3.3)$$

With this, the Average Black Run Length (ABRL) and Average White Black Transition count (AWBTC) of the content blocks has been computed using the BRL and WBTC values of every row (Mean of all BRL and WBTC). It is observed that the black run length appears to be more and the transition count appears low for the image regions. The ABRL produces a greater ratio (experimentally threshold has been fixed as greater than 0.4) for the image regions rather than the text regions. In contrast, the ABRL of the text regions appears to be lower than that of the image regions since the black run length is low while the transition count appears to be more. Apart from this, the AWBTC appears to be more for the text regions than for the images due to the large number of transitions in the text regions.

As a result, it is concluded that if the average black run length appears to be less (lesser than threshold-experimentally fixed and results are evaluated), and the average white black transition count appears to be more (AWBTC corresponds to 15-40% of total area for the text whereas it lies within 5-15% of the total area for the images), it is a text and vice versa for the images.

The results of the Text/Image separation system has been depicted in Figure 3.2, which show the image region in a different color. Details of the experiments conducted, data collected and the experimental evaluation of the RWSA system along with the text/image separation scheme is discussed in the following subsections.

III. EVALUATION OF SEGMENTATION METHODS

The performance of the proposed segmentation model is evaluated by the use of the measures such as Probabilistic Rand Index (Pantofaru and Hebert [29]), Variation of Information (Rubner et al., [30]), Global Consistency Error (Rubner et al., 2000) and Boundary Displacement Error (Schmid [31]).

A. Evaluation of Segmentation Approaches

Evaluation results vary significantly between different evaluators, because each evaluator may have distinct standards for measuring the quality of the segmentation.

Rand Index

Consider two images, say ground truth and segmented respectively: S1 and S2 of N points $X = \{x_1, x_2, x_3, ..., x_N\}$; that assigned labels $\{l_i\}$ and $\{l_i^{'}\}$ respectively to point x_i. The Rand Index can be computed as the ratio of the number of pairs of vertices having the compatible label relationship in S1 and S2. It can be defined as:

$$R(S_1, S_2) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i \neq j}} \left\lfloor \left\lceil I\left(l_i = l_j \wedge l_i^{'} = l_j^{'}\right) + I\left(l_i \neq l_j \wedge l_i^{'} \neq l_j^{'}\right) \right\rfloor \right\rceil \quad (1.1)$$

Where, I is the identity function, and the denominator is the number of possible unique pairs among N data points. This gives a measure of similarity ranging from 0 to 1.

Variation of Information

It measures the sum of information loss and information gain between the two clustering, and thus it roughly measures the extent to which one clustering can explain the other. For segmentations, it can be interpreted as the average conditional entropy of one segmentation given the other.

$$VI(S_{test}, S_K) = H(S_{test} \mid S_K) + H(S_K \mid S_{test}) \quad (1.2)$$

The first term in the above equation measures the amount of information about S_{test} that we lose, while the second term measures the amount of information about S_K that we have to gain, when going from segmentation S_{test} to ground truth S_K. Where, H (.|.) is the conditional entropy.

Global Consistency Error

Measures the extent to which the regions in one segmentation are subsets of the regions in second segmentation (i.e. the refinement). Let $R(S, p_i)$ be the set of pixels in segmentation S that contains pixel p_i, then the local refinement error is defined as:

$$E(S_1, S_2, p_i) = \frac{\left|R(S_1, p_i) \setminus R(S_2, p_i)\right|}{R(S_1, p_i)} \quad (1.3)$$

This error is not symmetric $\left(i.e., E(S_1, S_2, p_i) \neq E(S_2, S_1, p_i)\right)$ w.r.t. the compared segmentations, and takes the value of zero when S_1 is a refinement of S_2 at pixel p_i. Global Consistency Error is then defined as:

$$GCE(S_1, S_2) = \frac{1}{n} \min\left\{ \sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i) \right\} \quad (1.4)$$

where, n is the number of pixels.

Boundary Displacement Error (BDE)

The BDE is a boundary based metric to evaluate the segmentation quality. It defines the error of one boundary pixel as the distance between the pixel and its closest pixel in the other boundary image. Let *B*1, *B2* represent respectively the boundaries of segmentation and Ground truth. The BDE can be computed using the minimum absolute difference from arbitrary point *x* in *B*1 to all the boundary points in *B*2. A near-zero mean and small standard deviation of BDEs computed for all the points in *B*1 indicate the quality of the image segmentation.

i.e., $BDE = \min(x - y_i)$, where x∈B1 and $y_i \in B2$, $i = 1, 2, .., n$, n is the number of boundary points in B2.

IV. EXPERIMENTATION

To evaluate the segmentation results produced by different algorithms we have compiled a database, containing 300 corridor images along with ground truth segmentations. The corridor images was taken in twenty different buildings exhibiting a wide variety of different visual characteristic. Segmentation method is evaluated by assessing its consistency with the ground truth segmentation given by the human expert.

The segmentation is evaluated by assessing its consistency with the ground truth segmentation. Any evaluation metric desired should take into account the following effects: Over-segmentation where region of the reference is represented by two or more regions in the examined segmentation. Under segmentation were two or more regions of the reference are represented by a single region in the examined segmentation. In accurate boundary localization the ground truth is usually produced by humans that segment at different granularities. And finally in different number of segments one needs to compare two segmentations when they have different numbers of segments.

Table 1 shows the parameter values of different segmentation methods. The PRI value should be higher for an image and VOI, GCE. BDE values must be lower for an image. Each parameter is described by ground truth and proposed method. Each row is represented by average of each class totally about 100 images Form the table 1 the proposed method achieves values of PRI 0.9725, VI 2.23, GCE 2.14 and BDE 1.24, We can understand that proposed method achieves good results. From this evaluation, it is found that Region merging segmentation is well suited for the corridor images.

Table 1. Shows the segmentation results

Images no.	PRI		VI		GCE		BDE	
	Ground Truth	Proposed Method	Ground Truth	Proposed Method	Ground Truth	Proposed Method	Ground Truth	Proposed Method
1	0.9844	**0.9725**	0.9199	1.0986	2.4053	3.3784	0.2163	0.3480
2	0.9814	0.9668	0.8788	1.4768	3.2799	3.3368	0.2886	0.3087
3	0.9763	0.9766	0.9459	1.5755	**2.2390**	3.1061	0.2333	0.4429
4	0.9780	0.9699	0.8770	1.6447	2.4230	2.9160	0.2678	0.4030
5	0.9862	0.9702	0.8671	1.8866	3.1828	**2.5884**	0.3283	0.3421
6	0.9815	0.9665	0.8748	1.4235	2.4569	3.4137	0.2568	0.3388
7	0.9802	0.9652	0.8229	1.6247	2.9266	3.3850	0.2914	0.4066
8	0.9809	0.9694	0.8348	1.5332	2.7568	3.2784	0.2774	0.3917
9	0.9896	0.9633	0.9257	1.1464	2.7651	3.2690	0.1104	**0.2144**
10	0.9858	0.9669	0.9590	1.5422	3.2584	3.3537	0.1555	0.3495
11	0.9819	0.9628	0.9359	1.2741	2.6944	3.5492	0.2544	0.2803
12	0.9852	0.9631	**0.9575**	1.1895	3.2977	3.3552	0.1940	0.3639
13	0.9811	0.9615	0.8671	1.5554	2.7562	3.5547	0.3065	0.4117
14	0.9793	0.9625	0.9174	1.6896	2.5537	3.4230	0.3046	0.3925
15	0.9767	0.9693	0.9281	1.7558	2.6725	3.3752	0.2913	0.3946
16	0.9758	0.9628	0.9016	1.6596	2.4516	3.6604	0.2473	0.2882
17	0.9747	0.9631	0.9060	1.8337	2.6995	3.4166	0.2438	0.3454
18	0.9840	0.9615	0.9274	1.1838	2.7629	3.4586	0.2207	0.3583
19	0.9845	0.9625	0.8426	1.2308	2.3854	3.3339	0.2221	0.2488
20	0.9819	0.9693	0.8310	1.4975	2.7491	3.2971	0.2898	0.2541

Figure 3 Hybrid Layout Analysis Architecture

REFERENCES

[1] Tang Y. Y., Lee S. H and Suen C. Y., 1996. Automatic document processing: a survey. Pattern recognition, Vol. 29, No.12, pp. 1931-1952.

[2] Pal U. and Chaudhuri B.B. (2001), _Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line‘, Proceedings of the International Conference on Document Analysis and Recognition, pp. 790-794.

[3] Pal U., Sinha S. and Chaudhuri B.B. (2003), _Word-Wise Script Identification From A Document Containing English,Devnagari And Telugu Text‘, Proceedings of the Document Analysis and Recognition, pp. 213-220.

[4] Joshi G., Saurabh G. and Jayanthi S. (2006), _Script Identification from Indian Documents‘, Proceedings of the Seventh IAPR workshop on Document Analysis Systems, LNCS 3872, pp. 255-267.

[5] Busch A., Boles W.W. and Sridharan S. (2005), _Texture for Script Identification‘, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.11, pp.1720-1732.

[6] Spitz A.L. (1997), _Determination of script, language content of document images‘, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.3, pp. 235-245.

[7] Lu S. and Tan C.L. (2008), _Script and Language Identification in Noisy and Degraded Document Images‘, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 1, pp. 14-24.
[8] Pal U. and Chaudhuri B.B. (1997), _Automatic separation of words in multi-lingual multiscript Indian documents‘, Proceedings of the International Conference on Document Analysis and Recognition, pp. 576-579.
[9] Pal U. and Chaudhuri B.B. (1999), _Script Line Separation from Indian Multi-Script Documents,‘ Proceedings of the International Conference on Document Analysis and Recognition, pp.406- 409.
[10] Padma M.C. and Nagabhushan P. (2003), _Identification and Separation of text words of Kannada, Hindi and English languages through discriminating features‘, Proceedings of the National Conference of Document Analysis and Recognition, pp. 252-260.
[11] Dhandra B.V., Mallikarjun H., Ravindra H. and Malemath.V.S. (2007), _Word Level Script Identification in Bilingual Documents through Discriminating Features‘, Proceedings of International Conference on Signal processing, Communications and Networking, pp. 630-635.
[12] Dhanya D., Ramakrishnan A.G. and Peeta Basa P. (2002), _Script Identification In Printed Bilingual Documents,‘ Sadhana, Vol. 27, Part-1, pp. 73-82.
[13] Tan C.L., Leong P.Y. and He S. (1999), _Language Identification in Multilingual documents‘, Proceedings of the International Symposium on Intelligent Multimedia and Distance Education.
[14] Pal U. and Chaudhuri B.B. (1999), _Script Line Separation from Indian Multi-Script Documents,‘ Proceedings of the International Conference on Document Analysis and Recognition, pp.406- 409.
[15] Pal U., Sinha S. and Chaudhuri B.B. (2003), _Word-Wise Script Identification From A Document Containing English, Devnagari And Telugu Text‘, Proceedings of the Document Analysis and Recognition, pp. 213-220.
[16] Pati P.B., Sabari Raju S., Pati N. and Ramakrishnan A.G. (2004), _Gabor filters for document analysis in Indian Bilingual Documents‘, Proceedings of the International Conference on Intelligent Sensing and Information Processing, pp.123-126
[17] Pati P.B. and Ramakrishnan A.G. (2006), _HVS inspired system for script identification in Indian multi-script documents‘, Seventh IAPR Workshop on Document Analysis Systems, LNCS, Vol. 3872, pp. 380-389.
[18] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, Shet R.B., (2000), Identification of scripts ofIndian languages by Combining trainable classifiers‖, Proc. of ICVGIP, India.
[19] Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy, (2006), Script Identification from Indian Documents, DAS 2006, LNCS 3872, 255–267.
[20] Dhanya D., Ramakrishnan A.G. and Pati P.B., (2002), Script identification in printed bilingualdocuments, Sadhana, vol. 27, 73-82.
[21] Hiremath P S and S Shivashankar, ―Wavelet Based Co-occurrence Histogram Features for TextureClassification with an Application to Script Identification in a Document Image‖, PatternRecognition Letters 29, 2008, pp 1182-1189.
[22] Srinivas Rao Kunte R. and Sudhakar Samuel R.D., (2002), A Neural Approach in On-line ScriptRecognition for Telugu Language Employing Wavelet Features, National Workshop onComputer Vision, Graphics and Image Processing (WVGIP), 188-191.
[23] Peeta Basa Pati, S. Sabari Raju, Nishikanta Pati and A. G. Ramakrishnan, ―Gabor filters forDocument analysis in Indian Bilingual Documents‖, 0-7803-8243-9/04/ IEEE, ICISIP, pp. 123-126, 2004.
[24] Newsam, S. D., and Kamath, C.: Retrieval using texture features in high resolution multi-spectral satellite imagery. In SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology VI(2004).
[25] Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43(1):29-44, (2001).
[26] Schmid, C.: Constructing models for content-based image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 39-45(2001).
[27] Geusebroek, M., Smeulders, W. M., Weijer J.: Fast Anisotropic Gauss Filtering. IEEE Transactions on Image Processing, 12(8):938-943(2003).
[28] Varma, M., Zisserman A.: A statistical approach to texture classification from single images, International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis, 62(1--2):61—81(2005).
[29] Wasserman P D., Advanced Methods in Neural Computing, New York: Van Nostrand Reinhold (1993), pp. 155-61, and pp. 35-55(1993).
[30] Qasem, S. N., and Shamsuddin, S. M. :Generalization Improvement of Radial Basis Function Network Based on Multi- Objective Particle Swarm Optimization, Journal of Artificial Intelligence(2009).
[31] Lowe, David G. (1999). "Object recognition from local scale-invariant features". Proceedings of the International Conference on Computer Vision. pp. 1150–1157.
[32] Samet H., 1994. The Design and Analysis of Spatial Data Structures. Addison – Wesley.
[33] Chen J., H. Cao, R. Prasad, A. Bhardwaj and P. Natarajan, P., 2010. Gabor features for offline arabic handwriting recognition. Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems, pp. 53 – 58.
[34] Chen X. and J. Zhang, 2012. Optimized discriminant locality preserving projection of gabor feature for biometric recognition. International Journal of Security and Its Applications, vol. 6, no. 2, pp. 321 – 328.

International Journal of Computer Engineering and Applications,
Volume XII, Issue I, Jan. 18, www.ijcea.com ISSN 2321-3469

TEXTURAL FEATURES IN SCRIPT IDENTIFICATION FOR PRINTED BILINGUAL DOCUMENTS

Mahesha D M [1], Bhavya D N [2], Nandini H M [3]

[1]*Department of Studies in Computer Science, Karnataka State Open University, Mysore, India*
[2]*Department of Studies in Computer Science, Karnataka State Open University, Mysore, India*
[3]*Department of Studies in Information Technology, Karnataka State Open University, Mysore, India*

ABSTRACT:

In this work, we investigate the effect of texture features for script classification. Rectangular White Space analysis algorithm is used to analyze and identify heterogeneous layouts of document images. The texture features, namely the color texture moments, Local binary pattern (LBP) and responses of Gabor, LM-filter, S-filter, R-filter are extracted, and combinations of these are considered in the classification. In this work, a probabilistic neural network and Nearest Neighbor are used for classification. To corroborate the efficacy of the proposed method, an experiment was conducted on our own data set. The experiment was conducted for various sizes of the datasets, to study the effect of classification accuracy, and the results show that the combination of multiple features vastly improves the performance.

Keywords: Segmentation, Section finding, Section Merge, Feature Extraction, Classification.

[1] INTRODUCTION

Anything which conveys information is known as a document. Generally, a document is a knowledge container. Most of the times we acquire knowledge from documents such as Newspapers, Textbooks, Scientific journals, Magazines, Technical reports, Office files, Postal letters, Bank cheques, Application forms etc. (Tang et al., [1]). To understand the huge information, an extensive amount of manual processing is required and such a manual processing is very much time consuming. To overcome this difficulty, it is essential to automate the manual process which needs efficient algorithms. This automation process is considered as document

image processing (DIP). In general, the document image processing is divided into text processing and graphics processing. Text processing is further divided into character recognition and page layout analysis. Graphics processing is further divided into line processing and region processing as shown in Figure 1.

1.1 STAGES IN DOCUMENT IMAGE PROCESSING

The document image processing involves three basic steps at conceptual levels, which are document image analysis, document image recognition and document image understanding. Within these three levels, there are several other interacting modules such as image acquisition, binarization, block segmentation, block classification, logical block grouping, character and word recognition, picture processing and analysis, graphic analysis, picture understanding, text understanding and graphics understanding. The interactions between these processes and data flow between levels are shown in Figure 2.

1.1.1 DOCUMENT IMAGE ANALYSIS

Document image analysis is a process of recovering syntactic and semantic information from images of documents, prominently from scanned versions of paper documents. There are two distinct tasks in document image analysis. The first has a syntactical goal consisting of the identification of basic components of the document, the so-called document objects. The second has a semantic goal consisting of the identification of the role and meaning of the document objects in order to have an interpretation of the whole original document. The structural analysis, on the other hand involves usage of layout clues to identify headlines, locate different lines, etc. In general, image analysis involves the extraction and use of attributes and structure relationships in the document in order to label its components within contextual rules dictated by the document class. Analysis of printed documents obviously involves skew angle estimation and correction which is a very challenging task.

1.1.2 DOCUMENT IMAGE RECOGNITION

Document Image Recognition (DIR), a very useful technique in office automation and digital library applications, is to find the most similar template for any input document image in a prestored template document image. Nowadays a large amount of existing paper documents are transformed to digital document images through scanners and cameras. However, the next step is to analyze a document and segregate text blocks, graphic block, picture block, etc, so as to facilitate labeling of the blocks. This process of labeling the blocks is said to be document image recognition or identification.

1.1.3 DOCUMENT IMAGE UNDERSTANDING

Document image understanding is a component which extracts the logical relationships between the respective blocks of a document. Logical document structure is a hierarchical representation of semantics of the given document. The same logical document structure is formatted in varieties of physical layouts by changing the variables such as number of pages and font sizes, spacing between paragraphs and between sections, number of columns, etc. In all these layouts, the semantics of the document remains unaltered. Logical structure analysis determines the document’s semantic structure and provides data appropriate for information retrieval.

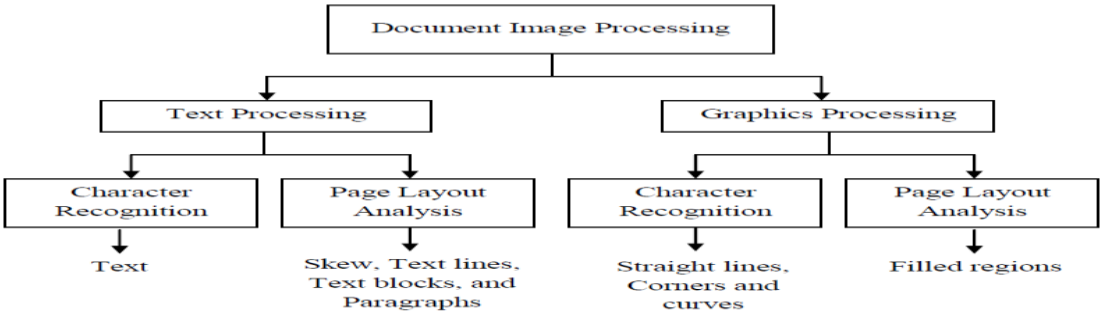


Figure 1. Hierarchy of document image processing with subcategories

1.2 SCRIPT RECOGNITION

The OCR technology for Indian documents is in emerging stage and most of these Indian OCR systems can read the documents written in only a single script. As per the Indian constitution, every state Government has to produce an official document containing a national language (Hindi), official language (English) and state language (or regional language). According to the two-language policy adopted by most of the Indian states, the documents produced in an Indian state Karnataka, are composed of texts in the regional language-Kannada and the world wide commonly used language-English. In addition, majority of the documents found in most of the private and Government sectors of Indian states, are bi-lingual type (a document having text in three languages). So, there is a growing demand to automatically process these bi-lingual documents in every state in India, including Karnataka.

The monolingual OCR systems will not process such multi-script documents without human involvement for delineating different script zones of multi-lingual pages before activating the script specific OCR engine. The need for such manual involvement can result in greater expense and crucially delays the overall image-to-text conversion. Thus, an automatic forwarding is required for the incoming document images to handover this to the particular OCR engine depending on the knowledge of the intrinsic scripts. In view of this, identification of script and/ or language is one of the elementary tasks for multi-script document processing. A script recognizer, therefore, simplifies the task of OCR by enhancing the accuracy of recognition and reducing the computational complexity. Script Recognition approaches can be broadly classified into two categories, namely, local and global approaches. The local approaches(Pal and Chaudhury [2], Pal et al [3]) analyze a list of connected components (Line, word, char) in the document images, to identify the script(or class of script). In contrast, global approaches (Joshi [4]) employ an analysis of regions (block of text) comprising at least two lines (or words)without finer segmentation. In general, global approaches work well based on texture measurement, but this relies heavily on a uniform block of text (Buschet al [5]), and extensive preprocessing (to make the text block uniform) is required to measure the texture. Even though local approaches rely on the accuracy of character segmentation or connected component analysis, it could work well on the documents irrespective of their quality or uniformity in the block of text.

In the literature, many works have been reported for script recognition at the document, line and word levels, using local approaches. In this context, researchers have made a number of attempts to discriminate the Han and Latin script (Spitz [6], Lu and Tan [7]) at the document level and exploited many Indian scripts at line level and word level (Pal and Chaudhury [8], Pal and Chaudhury [9], Padma and Nagabhushan [10], Dhandra et al [11]). However, all the techniques reported in the literature are script dependent. Since this research is intended to develop an classification system for kannada document images, Script Recognition, to discriminate the kannada from English scripts in bilingual document images is becoming important. In this connection, few local approaches are reported in the literature, such as spatial spread analysis (Dhanya et al [12]), Aspect Ratio (Tan et al [13]), Structural features (Pal and Chaudhury [14]), and Water Reservoirs (Pal et al [15]).

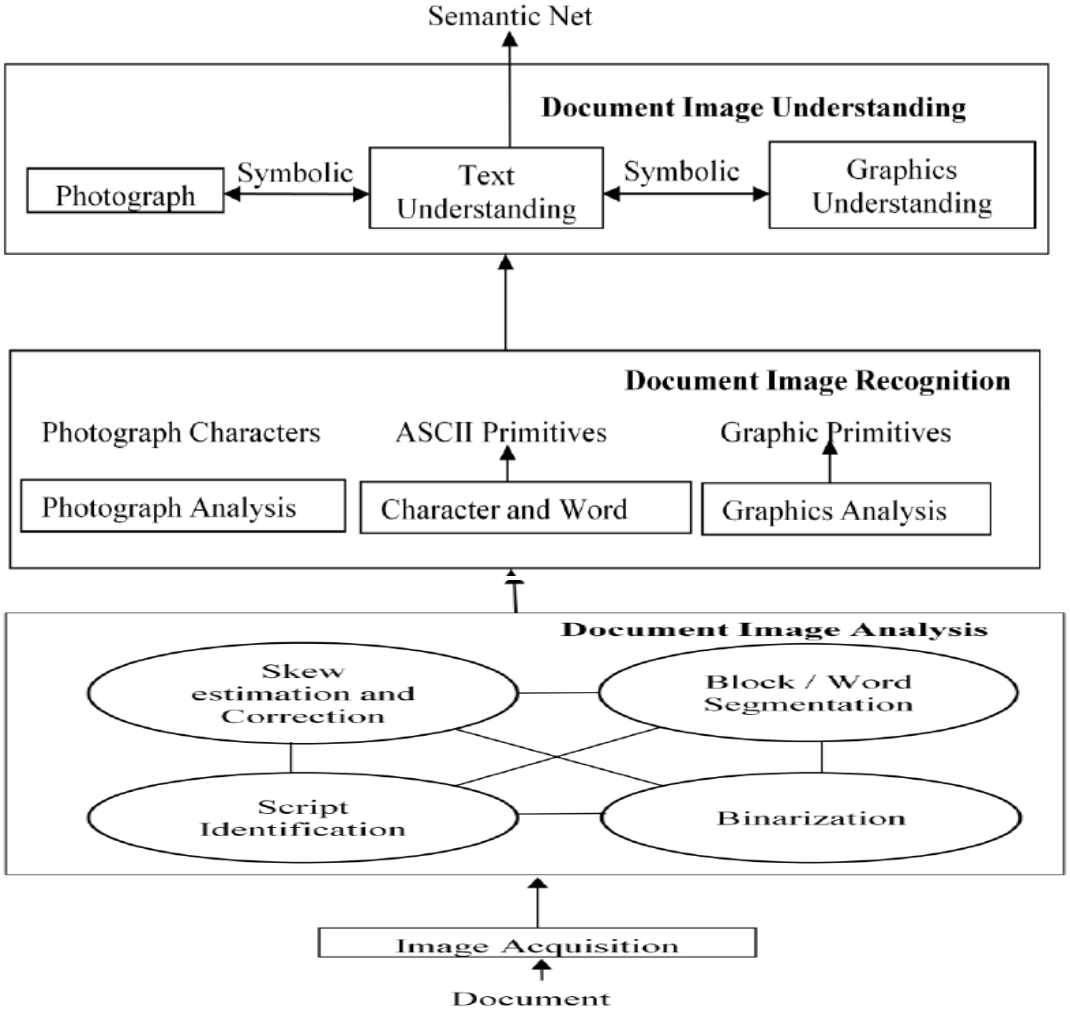


Figure 2. Steps involved in document image processing

However, all the above mentioned techniques produce a low discrimination rate due to its incapability in exploration of thescripts. Global approaches (Pati et al [17], Pati and Ramakrishnan [18]). S Chaudhury et al., [19] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. G D Joshi et al., [20] have presented a script identification technique for 10 Indian scripts using a set of features extracted

from logGabor filters. Dhanya et al., [21] have used Linear Support Vector Machine (LSVM), K-Nearest Neighbour (K-NN) and Neural Network (NN) classifiers on Gabor-based and zoning features to classify Tamil and English scripts. Hiremath [22] have proposed a novel approach for script identification of South Indian scripts using wavelet based co-occurrence histogram features. S R Kunte and S Samuel [23] have suggested a neural approach in on-line script recognition for Telugu language employing wavelet features. Peeta et al., [24] have presented a technique using Gabor filters for script identification of Indian bilingual documents.

[2] SEGMENTATION

The preprocessing of the document images includes the process of Noise removal and Binarization. Noise could occur in the document images due to many sources such as aging, photocopying etc. and the application of filters reduces noises in these images. Here noise has been suppressed in the document image by using a median filter, since median filters smear the character image strokes. For this median filter, a 3*3 mask has been chosen and it is applied over the image, which replaces nine pixels by the intensity of the center pixel over this mask. As a result of pulling the median filter output to the gray level of the center pixel, the shapes of the character strokes can be preserved. Binarization has been applied after noise removal. Binarization is a technique by which the color and gray scale images are converted into binary images. The most common method is to select a proper threshold for the image and convert all the intensity values above the threshold into an intensity value representing as ‘white’ and below the threshold as ‘black’ value. All intensity values below a threshold are converted to one intensity level and intensities higher than this threshold are converted to the other chosen intensity.

Since all the books and magazines use white spaces as a separator within and between the texts, the observation of small white spaces becomes mandatory to identify the text area. Therefore, in Section Finding, white spaces are used as delimiters and observed for analysis. Variable length white spaces exist inside the text in both the directions, apart from the white spaces surrounding the textual zones. Due to the existence of non-uniform, small white gaps in the image apart from the column separators, a careful analysis is required to observe and record the white spaces. As a result, the width of the image has been divided into ‘n’ equal sections. Since connected component analysis has been eliminated, a single horizontal scan has been performed over the image to grab the white spaces. After an entire horizontal scan of an image, all the sections which appear as white spaces are reported and their positions with the corresponding row number have been recorded as a result of this procedure. It is hard to process various white space section numbers to identify the layout gaps if the merging procedure has been avoided. Once all the white space section numbers based on number have been indicated, the merging of adjacent sections in both the directions is required to form horizontal and vertical white space rectangles which are done through the Section Merging phase.

The Section Merging phase consists of two processes: Horizontal Section Merging and Vertical Section Merging. Initially, horizontal section merging accepts all the white space section numbers with their corresponding row numbers as the input and produces a series of within-line or row-wise white space clusters as output(i.e.), subsequent white space sections in each row gets merged together to produce a series of row-wise white space sections. Since all the white spaces (section-wise) are identified and merged properly, the chance of getting under-segmentation has

been completely eliminated. The rectangular analysis phase consists of Cropping and the Rectangular formation process. After the identification of horizontal and vertical white space rectangles, finding the areas which are uncovered by the white space rectangles could yield the layout. The Cropping procedure acts over the white space rectangles in both the directions by accepting the horizontal edges of each Horizontal White Space Rectangle (HWSR) and the vertical edge of each Vertical White Space Rectangle (VWSR). Once the horizontal and vertical edges are cropped, the areas uncovered by the white spaces could be easily extracted through rectangular formation procedure. Once the content blocks have been identified, the next step attempts to separate the textual blocks from the images and pictures, since textual blocks are required for further processing. Once the homogeneous regions are obtained, each region gets passed into the text image analyzer to identify the text component. Two statistical properties called as Black Run Length (BRL) and White Black Transition Count (WBTC), which spans in the horizontal direction of the image have been used here to identify the textual blocks. Black run length corresponds to the ratio of the total number of black pixels in a row to the total transition (black-white disposition) count in that row. The White Black Transition count corresponds to the ratio of the total number of transitions in a row to the total number of pixels in that row. It is concluded that if the mean black run length appears to be more, and the Mean white black transition count of all the rows appears to be lesser than the threshold, it is concluded as image and not as a text.

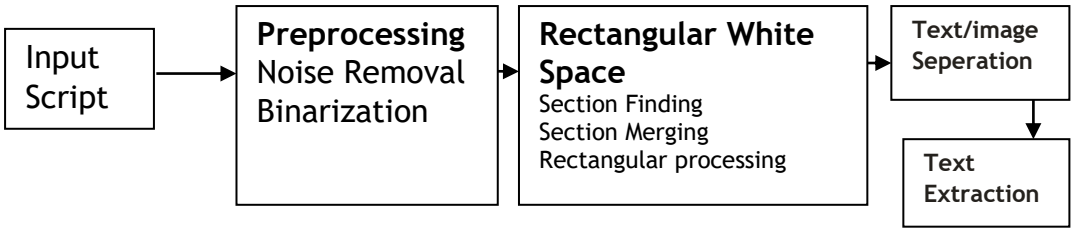


Figure 3. Steps involved in text extraction

Figure 4. Input Image with text

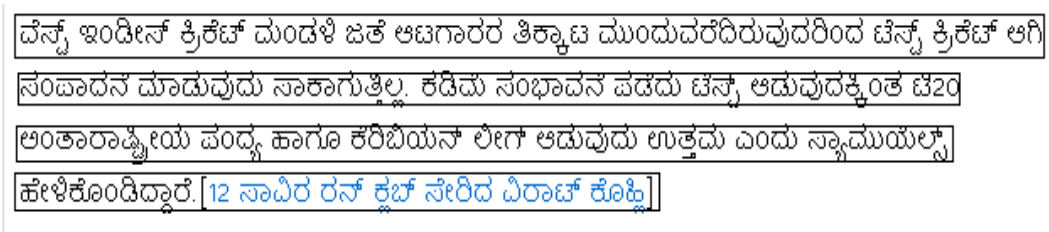


Figure 5. Input Image with elimination text

[3] FEATURE EXTRACTION

As our interest is to study only texture features for script classification, for a segmented image we extract Color Texture Moments, Local binary pattern, LM filter responses, Schmid filter responses, Maximum filter responses and Gabor responses for texture analysis. The flowing subsection gives an introduction to all the above texture features.

3.1 GABOR FILTER RESPONSES

The Texture analysis using filters based on Gabor functions falls into the category of frequency-based approaches. These approaches are based on the premise that texture is an image pattern containing a repetitive structure that can be effectively characterized in a frequency domain, such as the Fourier domain. One of the challenges, however, of such an approach is dealing with the tradeoff between the joint uncertainty in the space and frequency domains. Meaningful frequency based analysis cannot be localized without bound. An attractive mathematical property of Gabor functions is that they minimize the joint uncertainty in space and frequency. They achieve the optimal tradeoff between localizing the analysis in the spatial and frequency domains. Using Gabor filters to analyze texture appeals from a psycho-visual perspective as well. The texture analysis is accomplished by applying a bank of scale and orientation selective Gabor filters to an image [25]. These filters are constructed as follows. A two-dimensional Gabor function g(x; y) and its Fourier transform G (u; v) can be written as:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^{\,2}} + \frac{y^2}{\sigma_y^{\,2}}\right) + 2\pi i W x\right]$$

and

$$G(u,v) = \exp\left\{-\frac{1}{2}\left[\frac{(u-W)^2}{\sigma_u^{\,2}} + \frac{v^2}{\sigma_v^{\,2}}\right]\right\}$$

Where $i = \sqrt{-1}$, $\sigma_u = 1/2\pi\sigma_x$ and $\sigma_v = 1/2\pi\sigma_y$ control the tradeoff between spatial and frequency resolution, and W controls the modulation. A class of self-similar functions referred to as Gabor wavelets is now considered. Let g(x, y) be the mother wavelet. A filter dictionary can be obtained by appropriate dilations and translations of g(x, y) through the generating function.

$$g_{rs}(x,y) = a^{-s} g(x',y'),\ a > 1, s\varepsilon 0,1..S-1, r\varepsilon 1,2,..R$$

$x' = a^{-s}(x\cos\theta + y\sin\theta)$ and $y' = a^{-s}(-x\sin\theta + y\cos\theta)$

where $\theta = (r-1)\pi / R$. The indices r and s indicate the orientation and scale of the filter respectively. R is the total number of orientations and S is the total number of scales in the filter bank. While the size of the filter bank is application dependent, experimentation has shown that a bank of filters tuned to combinations of 0, 2, 4, 6, 8, 10 scales, and different orientations, at 22.5 degree intervals is sufficient for flower analyses.

3.2 MAXIMUM RESPONSE (MR) FILTER RESPONSES

We describe the texture by convolving the images with MR filter bank introduced by Varma and Zisserman[26]. The MR sets contain both isotropic filters as well as anisotropic filters at multiple orientations they generate good features for all types of textures. Additionally, unlike traditional rotationally invariant filters, the MR sets are also able to record the angle of maximum response. This enables to compute higher order co-occurrence statistics on orientation and such statistics may prove useful in discriminating textures which appear to be very similar. The MR filter banks generate more significant textons not only because of improved clustering in a lower dimensional space but also because rotated features are correctly mapped to the same texton.

3.3 LEUNG-MALIK (LM) AND SCHMID FILTER RESPONSES

The LM set is a multi scale, multi orientation filter bank with 48 filters. It consists of first and second derivatives of Gaussians at 6 orientations and 3 scales making a total of 36; 8 Laplacian of Gaussian (LOG) filters; and 4 Gaussians. In [27] two versions of the LM filter bank are considered. In LM Small (LMS), the filters occur at basic scales $\sigma = \{1, \sqrt{2}, 2, 2\sqrt{2}\}$. The first and second derivative filters occur at the first three scales with an elongation factor of 3 (i.e. $\sigma_x = \sigma$ and $\sigma_y = 3\sigma_x$). The Gaussians occur at the four basic scales while the 8 LOG filters occur at σ and 3σ. For LM Large (LML), the filters occur at the basic scales $\sigma = \{\sqrt{2}, 2, 2\sqrt{2}, 4\}$. A Schmid response [14] is rotational invariant but the invariance is achieved in a different manner and texton clustering is always in a higher dimensional space. The Schmid set consists of 13 rotationally invariant filters of the form.

$$F(r,\sigma,\tau) = F_0(\sigma,\tau) + \cos\left(\frac{\pi\tau r}{\sigma}\right) e^{-\frac{r^2}{2\sigma^2}}$$

3.4 COLOR TEXTURE MOMENTS (CTM)

Yu et al., in their work [28] developed a color and texture combined moments for image retrieval. In their work the image was converted onto different color space viz., RGB, HSV, YUV, and (SVcosH, SVsinH, V). Local Fourier Transform (LFT) was performed on all the channels of image with eight different templates and for each resulted channel image first two moments were calculated obtaining 48 features for a given single image. Experimentally they found out that (SVcosH, SVsinH, V) color space is better than other color space and these features were termed as color texture moments. In our work we have used this color texture moments as one of the features.

3.5 LOCAL BINARY PATTERN

Ojala et al. [29] proposed to use the Local Binary Pattern (LBP) histogram for rotation invariant texture classification. LBP is a simple but efficient operator to describe local image patterns. It is combined statistical and structured method.LBP is a gray-scale texture operator that characterizes the local spatial structure of the image texture. The basic LBP operator considers a 3x3 neighborhood of a pixel, then these 8 border pixels will be replaced either by 1, if they are larger than or equal to the central pixel or by 0 otherwise. Finally, the central pixel will be replaced with a summation of the binary weights of border pixels in the LBP image and the 3x3 window slides to the next pixel. It is possible to develop the basic LBP into various neighborhood sizes and distances.

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(g_P - g_c)2^P \quad (1)$$

Where s (.) is the sign function:

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2)$$

$g_p$and $g_c$are grey levels of border pixels and central pixel respectively, P is the number of pixels in the neighborhood and R is the radius of the neighborhood. Suppose the coordinates of g_c are (0, 0), then the coordinates of g_p are given by (-R sin (2πp/P), R cos (2πp/P)). In this case, if we set (P = 8; R = 1), we obtain the basic LBP (1) Luminance changing cannot affect signed differences g_p-g_c, hence LBP is grey level shift invariant.

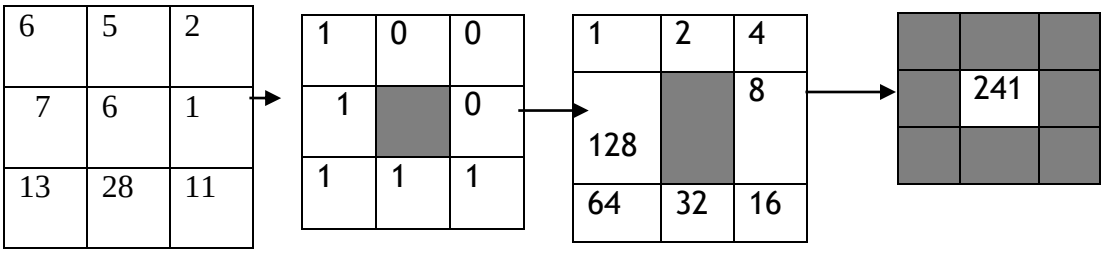


Image Threshold Weights LBP=1+16+32+64+128=241

Fig 6: Computing the basic 3x3 LBP.

Suppose the texture image is NxM. After identifying the LBP pattern of each pixel (i, j), the whole texture image is represented by building a histogram:

$$H(k) = \sum_{i=1}^{N} \sum_{j=1}^{M} f(LBP_{P,R}(i,j), k), k \in [0, K] \quad (3)$$

$$f(x, y) = \begin{cases} 1, & x = y \\ 0, & otherwise \end{cases} \quad (4)$$

where K is the maximal LBP pattern value. The U value of an LBP pattern is defined as the number of spatial transitions (bitwise0/1 changes) in that pattern

$$U\left(LBP_{P,R}\right) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{P=1}^{P-1} |s(g_P - g_c) - s(g_{P-1} - g_c)| \quad (5)$$

For example LBP pattern 00000000 has a U value of 0 and 01000000 of 2.The uniform LBP pattern refers to the uniform appearance pattern which has limited transition or discontinuities (U≤2) in the circular binary presentation. It was verified hat only ‘‘uniform’’ patterns are fundamental patterns of local image texture.In practice, the mapping from $LBP_{P,R}$to $LBP^{u2}_{P,R}$ (superscript “*u2*” means uniform patterns with U≤2),which has P*(P-1)+3 distinct output values, is implemented with a lookup table of 2^Pelements. To achieve rotation invariance, a locally rotation invariant pattern could be defined as:

$$LBP_{P,R}^{riu^2} = \begin{cases} \sum_{P=0}^{P-1} s(g_P - g_c) & if\ U\left(LBP_{P,R}\right) \leq 2 \\ P + 1 & otherwise \end{cases} \quad (6)$$

The mapping from $LBP_{P,R}$to $LBP_{P,R}^{riu^2}$(superscript “*riu2*”means rotation invariant “uniform” patterns with U≤2), which has P+2 distinct output values, can be implemented with a lookup table.

[4] CLASSIFICATION

The extracted texture features is fed to fusion classifier of nearest neighbor and Probabilistic Neural Network. Introductions on each of these classifiers are given in the flowing subsections.

4.1 NEAREST NEIGHBOR (KNN)

One of the simplest classifiers which we used is the K-Nearest Neighbor classifier [29] [30]. The term nearest can be taken to mean the smallest Euclidean distances in n-dimensional feature space. This takes a test sample feature in a vector form, and finds the Euclidean distance between this and the vector representation of each training example. The training sample closest to the test sample is termed as its Nearest Neighbor. Since the trained sample in some sense is the one most similar to our test sample, it makes sense to allocate its class label to the test sample. This exploits the ‘smoothness’ assumption that samples near each other are likely to have the same class.

4.2 PROBABILISTIC NEURAL NETWORKS (PNN)

Probabilistic neural networks ([29] [30]) are forward feed networks built with three layers. They are derived from Bayes Decision Networks. They train quickly since the training is done in one pass of each training vector, rather than several. Probabilistic neural networks estimate the probability density function for each class based on the training samples. The probabilistic neural network uses a similar probability density function. This is calculated for each test vector. Vectors must be normalized prior to input into the network. There is an input unit for each dimension in the vector. The input layer is fully connected to the hidden layer. The hidden layer

has a node for each classification. Each hidden node calculates the dot product of the input vector with a test vector subtracts 1 from it and divides the result by the standard deviation squared. The output layer has a node for each pattern classification. The sum for each hidden node is sent to the output layer and the highest values wins. The Probabilistic neural network trains immediately but execution time is slow and it requires a large amount of space in memory.

4.3 FUSION APPROACH

In this work we study the performance of each classifier individually. Then we also investigate the suitability of combining two classifiers together. The proposed fusion is a decision level fusion approach using OR operation.

[5] RESULTS

In this experimentation, we intend to study the performances of different texture features in script classification and also effect of classifiers. The experimentation has been conducted on database of 30 classes under varying training samples from 40, 60 and 80 percent of database. The results obtained for individual feature, Combination of single feature, Combination of two feature, Combination of three feature and Combination of four feature are respectively tabulated in Table 1, Table2, Table 3,Table 4 and Table 5. In each table, the results are tabulated for each individual classifier and fusion of two classifiers. It shall be noticed that the fusion of all classifiers achieves relatively higher accuracy in all cases.

Features	KNN	PNN	Fusion
Gabor	61.11	**57.78**	71.11
	44.76	42.86	56.67
	32.12	30	39.7
CTM	**80**	22.22	**82.22**
	60	19.05	63.81
	36.36	15.76	41.82
LMFilter	17.78	15.56	32.22
	15.24	16.19	29.05
	10.91	10	19.09
Sfilter	18.89	16.67	36.67
	15.24	12.73	29.05
	11.21	8.1	22.42
Rfilter	16.19	10.95	22.22
	13.33	9.09	26.19
	12.22	7.78	20
LBP	18.89	9.09	31.11
	15.24	8.57	24.76
	13.64	7.78	21.21

Table 1: Accuracy of different classifier and their combinations for individual Texture Features

Features	KNN	PNN	Fusion
Gabor+CTM	62.22	60	**68.89**
	48.1	45.24	60.95
	33.94	30.61	43.94
Gabor+LMfilter	42.22	32.22	54.44
	33.81	23.81	47.14
	21.82	16.06	32.42
Gabor+Sfilter	52.22	38.89	62.22
	39.05	26.19	51.43
	25.45	15.15	33.94
Gabor+Rfilter	41.11	34.44	50
	29.05	22.38	44.76
	20.61	15.15	33.64
Gabor+LBP	56.67	57.78	72.22
	45.24	43.81	59.52
	32.12	29.39	40.91
CTM+LmFilter	21.11	17.78	40
	20.95	15.24	35.71
	14.55	12.73	25.15
CTM+Sfilter	27.78	20	47.78
	19.05	15.24	32.86
	13.33	13.33	24.24
CTM+Rfilter	22.22	18.89	35.56
	18.1	13.33	30.48
	15.15	11.21	26.06
CTM+LBP	26.67	22.22	42.22
	22.38	23.33	38.1
	20.3	17.58	30.3
LMfilter+Sfilter	18.89	13.33	33.33
	14.76	12.38	23.33
	7.58	6.36	17.27
LMfilter+Rfilter	13.33	11.11	24.44
	13.81	5.15	22.38
	7.88	9.52	17.58
Lmfilter+LBP	20	15.56	31.11
	14.29	12.86	24.29
	12.12	10.61	22.12
Sfilter+Rfilter	13.33	12.22	22.22
	11.9	9.52	20.95
	9.39	8.18	17.27

Sfilter+LBP	18.89	14.44	33.33
	10.95	10.95	22.38
	9.39	10	19.7
Rfilter+LBP	16.67	12.73	29.52
	12.73	10.48	24.24
	12.22	10	23.33

Table 2: Accuracy of different classifier and their combinations for Combination Two features

Features	KNN	PNN	Fusion
Gabor+CTM+LMfilter	42.22	31.11	53.33
	34.29	25.24	47.14
	23.03	17.27	35.15
Gabor+CTM+Sfilter	53.33	40	65.56
	39.52	28.1	52.86
	26.36	16.06	34.85
Gabor+CTM+Rfilter	43.33	34.44	52.22
	31.43	22.86	47.62
	21.52	15.76	33.03
Gabor+CTM+LBP	57.78	60	**71.11**
	47.14	46.67	60
	33.33	31.21	43.03
CTM+LMfilter+Sfilter	21.11	22.22	37.78
	18.1	17.62	31.9
	8.48	12.73	19.09
CTM+Sfilter+Rfilter	17.78	16.67	32.22
	16.19	11.43	31.9
	10.3	8.79	23.33
CTM+Rfilter+LMfilter	16.67	13.33	27.78
	14.76	12.86	27.62
	7.88	6.36	23.03
Gabor+Rfilter+LMfilter	25.56	15.56	45.55
	23.81	15.24	35.71
	16.36	10.61	27.27
Gabor+Rfilter+Sfilter	32.22	18.88	47.78
	25.24	14.76	42.38
	16.97	9.09	27.58
Gabor+Sfilter+LMfilter	34.44	18.89	47.78
	27.62	19.05	41.43
	18.18	10.61	30.3
Gabor+Rfilter+LBP	42.22	34.44	52.22
	29.52	21.9	45.24

	21.21	14.85	32.12
Gabor+LMfilter+LBP	41.11	32.22	54.44
	31.43	22.38	46.19
	22.42	15.76	33.33
Gabor+Sfilter+LBP	55.56	36.67	66.67
	37.14	25.71	52.38
	24.55	16.06	36.36
CTM+Rfilter+LBP	22.22	17.78	37.78
	21.43	15.24	33.81
	16.36	13.33	29.09
CTM+LMfilter+LBP	24.44	18.89	36.67
	20.95	17.62	36.19
	13.94	14.24	25.45
CTM+Sfilter+LBP	18.89	15.56	42.22
	12.38	13.81	28.57
	12.42	11.21	26.97
Rfilter+Sfilter+LBP	14.44	12.22	25.56
	12.86	9.52	21.9
	10.91	8.48	21.21
Rfilter+LMfilter+LBP	13.33	11.11	31.11
	11.43	9.52	21.43
	6.67	5.45	16.06
Sfilter+LMfilter+LBP	18.89	13.33	30
	15.24	12.86	28.1
	7.58	6.97	17.27

Table 3: Accuracy of different classifier and their combinations for Combination of Three features

Gabor+CTM+LMfilter+Rfilter	30	24.44	45.56
	25.71	22.86	37.62
	16.36	18.79	28.48
Gabor+CTM+LMfilter+Sfilter	36.67	25.56	50
	29.52	27.14	44.76
	19.39	19.09	32.42
Gabor+LBP+LMfilter+Sfilter	36.67	27.78	55.56
	28.1	24.76	42.86
	20.3	15.45	31.21
Gabor+LBP+LMfilter+Rfilter	25.56	30	47.78
	23.81	23.81	36.67
	16.97	20.3	30.91
Gabor+LBP+Sfilter+Rfilter	32.22	30	45.56
	25.71	26.67	42.38
	17.88	17.88	27.88

Gabor+LBP+CTM+Rfilter	44.44	33.33	54.44
	30.95	22.38	46.19
	22.42	15.15	34.55
Gabor+LBP+CTM+Sfilter	44.44	37.78	**70**
	30.95	27.14	53.33
	22.42	16.36	38.18
Gabor+LBP+CTM+LMfilter	42.22	32.22	58.89
	32.86	24.29	47.14
	23.33	17.58	33.94
Gabor+CTM+Rfilter+Sfilter	35.56	21.11	50
	27.62	15.71	45.24
	18.79	11.21	31.21
Gabor+LMfilter+Rfilter+Sfilter	23.33	25.56	37.78
	23.81	20.48	34.29
	14.24	16.67	25.76
CTM+LMfilter+Rfilter+Sfilter	23.81	11.11	24.44
	16.19	9.05	23.81
	14.55	6.36	19.7
LBP+LMfilter+Rfilter+Sfilter	16.67	8.89	26.67
	13.03	7.62	20.95
	11.9	5.76	18.48
LBP+CTM+ Rfilter+Sfilter	18.89	17.78	35.56
	12.38	11.9	30.48
	10.91	9.09	24.55
LBP+CTM+ Sfilter+LMfilter	24.44	24.44	42.22
	17.62	17.14	31.9
	8.48	12.42	19.39
LBP+CTM+ Rfilter+LMfilter	20	11.11	27.78
	17.62	10.95	27.14
	16.06	6.36	23.03

Table 4: Accuracy of different classifier and their combinations for Combination of Four features

Features	KNN	PNN	Fusion
Gabor+CTM+LMfilter+Rfilter+Sfilter	24.44	25.56	41.11
	24.29	23.81	37.62
	14.24	16.67	26.36
Gabor+LBP+LMfilter+Rfilter+Sfilter	24.44	25.56	40
	23.33	20.95	34.29
	13.94	16.06	25.76

Gabor+CTM+LBP+Rfilter+Sfilter	32.22	28.89	45.56
	27.14	27.14	43.81
	18.79	20.61	31.21
Gabor+CTM+LBP+LMfilter+Rfilter	28.89	25.56	45.56
	24.76	21.43	35.24
	17.58	20.3	30
Gabor+CTM+LBP+LMfilter+Sfilter	37.78	28.89	**55.56**
	29.52	28.57	46.19
	20.91	18.79	34.24
LMfilter+CTM+LBP+Rfilter+Sfilter	19.52	11.11	27.78
	18.89	9.52	26.67
	16.06	6.36	20.3

Table 5: Accuracy of different classifier and their combinations for Combination of Five features

[6] CONCLUSION

In this work we develop a bi-lingual script classification system based on the combination of texture features and classifiers. The suitable texture features such as Color Texture Moments, Local binary pattern, LM filter responses, Schmid filter responses, Maximum filter responses and Gabor responses is explored for the purpose of plant classification. It is observed that the fusion classifier achieves relatively good classification accuracy when compared to any other available classifier. We have created our own database of documents. We conducted experimentation under varying database size and we studied its effect on classification accuracy. The experimental results have shown that the fusion classifier outperforms any individual classifier.

REFERENCES

1. Tang Y. Y., Lee S. H and Suen C. Y., 1996. Automatic document processing: a survey. Pattern recognition, Vol. 29, No. 12, pp. 1931-1952.
2. Pal U. and Chaudhuri B.B. (2001), ‘Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line’, Proceedings of the International Conference on Document Analysis and Recognition, pp. 790-794.
3. Pal U., Sinha S. and Chaudhuri B.B. (2003), ‘Word-Wise Script Identification From A Document Containing English, Devnagari And Telugu Text’, Proceedings of the Document Analysis and Recognition, pp. 213-220.
4. Joshi G., Saurabh G. and Jayanthi S. (2006), ‘Script Identification from Indian Documents’, Proceedings of the Seventh IAPR workshop on Document Analysis Systems, LNCS 3872, pp. 255-267.
5. Busch A., Boles W.W. and Sridharan S. (2005), ‘Texture for Script Identification’, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.11, pp.1720-1732.
6. Spitz A.L. (1997), ‘Determination of script, language content of document images’, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.3, pp. 235-245.

7. Lu S. and Tan C.L. (2008), 'Script and Language Identification in Noisy and Degraded Document Images', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 1, pp. 14-24.
8. Pal U. and Chaudhuri B.B. (1997), 'Automatic separation of words in multi-lingual multiscript Indian documents', Proceedings of the International Conference on Document Analysis and Recognition, pp. 576-579.
9. Pal U. and Chaudhuri B.B. (1999), 'Script Line Separation from Indian Multi-Script Documents,' Proceedings of the International Conference on Document Analysis and Recognition, pp.406- 409.
10. Padma M.C. and Nagabhushan P. (2003), 'Identification and Separation of text words of Kannada, Hindi and English languages through discriminating features', Proceedings of the National Conference of Document Analysis and Recognition, pp. 252-260.
11. Dhandra B.V., Mallikarjun H., Ravindra H. and Malemath.V.S. (2007), 'Word Level Script Identification in Bilingual Documents through Discriminating Features', Proceedings of International Conference on Signal processing, Communications and Networking, pp. 630-635.
12. Dhanya D., Ramakrishnan A.G. and Peeta Basa P. (2002), 'Script Identification In Printed Bilingual Documents,' Sadhana, Vol. 27, Part-1, pp. 73-82.
13. Tan C.L., Leong P.Y. and He S. (1999), 'Language Identification in Multilingual documents', Proceedings of the International Symposium on Intelligent Multimedia and Distance Education.
14. Pal U. and Chaudhuri B.B. (1999), 'Script Line Separation from Indian Multi-Script Documents,' Proceedings of the International Conference on Document Analysis and Recognition, pp.406- 409.
15. Pal U., Sinha S. and Chaudhuri B.B. (2003), 'Word-Wise Script Identification From A Document Containing English, Devnagari And Telugu Text', Proceedings of the Document Analysis and Recognition, pp. 213-220.
16. Pati P.B., Sabari Raju S., Pati N. and Ramakrishnan A.G. (2004), 'Gabor filters for document analysis in Indian Bilingual Documents', Proceedings of the International Conference on Intelligent Sensing and Information Processing, pp.123-126
17. Pati P.B. and Ramakrishnan A.G. (2006), 'HVS inspired system for script identification in Indian multi-script documents', Seventh IAPR Workshop on Document Analysis Systems, LNCS, Vol. 3872, pp. 380-389.
18. Santanu Chaudhury, Gaurav Harit, Shekar Madnani, Shet R.B., (2000), Identification of scripts ofIndian languages by Combining trainable classifiers", Proc. of ICVGIP, India.
19. Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy, (2006), Script Identification from IndianDocuments, H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, 255–267.
20. Dhanya D., Ramakrishnan A.G. and Pati P.B., (2002), Script identification in printed bilingual documents, Sadhana, vol. 27, 73-82.
21. Hiremath P S and S Shivashankar, "Wavelet Based Co-occurrence Histogram Features for Texture Classification with an Application to Script Identification in a Document Image", Pattern Recognition Letters 29, 2008, pp 1182-1189.
22. Srinivas Rao Kunte R. and Sudhakar Samuel R.D., (2002), A Neural Approach in On-line Script Recognition for Telugu Language Employing Wavelet Features, National Workshop on Computer Vision, Graphics and Image Processing (WVGIP), 188-191.
23. Peeta Basa Pati, S. Sabari Raju, Nishikanta Pati and A. G. Ramakrishnan, "Gabor filters for Document analysis in Indian Bilingual Documents", 0-7803-8243-9/04/ IEEE, IClSlP, pp. 123-126, 2004.
24. Newsam, S. D., and Kamath, C.: Retrieval using texture features in high resolution multi-spectral satellite imagery. In SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology VI(2004).
25. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43(1):29-44, (2001).
26. Schmid, C.: Constructing models for content-based image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 39-45(2001).
27. Geusebroek, M., Smeulders, W. M., Weijer J.: Fast Anisotropic Gauss Filtering. IEEE Transactions on Image Processing, 12(8):938-943(2003).

28. Varma, M., Zisserman A.: A statistical approach to texture classification from single images, International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis, 62(1--2):61—81(2005).
29. Wasserman P D., Advanced Methods in Neural Computing, New York: Van Nostrand Reinhold (1993), pp. 155-61, and pp. 35-55(1993).
30. Qasem, S. N., and Shamsuddin, S. M. :Generalization Improvement of Radial Basis Function Network Based on Multi-Objective Particle Swarm Optimization, Journal of Artificial Intelligence(2009).

Abrupt Shot Change Detection using Midhinge Local Binary Pattern

Nandini H M [1, 2]
[1] Department of Computer Science and Engineering, Maharaja Research Foundation, Maharaja Institute of Technology, Mysuru, India
[2] Department of Information Technology, Karnataka State Open University, Mysuru, India
nandiniksou@gmail.com

Chethan H K [1]
[1] Department of Computer Science and Engineering, Maharaja Research Foundation, Maharaja Institute of Technology, Mysuru, India
hkchethan@gmail.com

Rashmi B S [2]
[2] Department of Information Technology, Karnataka State Open University, Mysuru, India
rashmibsrsh@compsci.uni-mysore.ac.in

***Abstract*— Video shot change detection is an extremely important task for video analysis and is considered as the preliminary step for effective video summarization, browsing and retrieval. In this work, a new texture based descriptor has been proposed using Midhinge Local Binary Pattern (MHLBP) concept for identification of abrupt transition in videos. The discrimination capability of MHLBP is better than basic LBP and its versions (mean LBP and midrange LBP). MHLBP histogram is constructed by applying midhinge statistics on each mask of video frame. The distance between histogram features of the adjacent frames are evaluated using Euclidean distance. The obtained distance values are subjected to adaptive thresh-old mechanism for identifying of abrupt shot changes in a video. The planned framework is tested on a subset of TRECVID 2001 dataset. The result shows that the proposed method outperforms with other existing shot change detection algorithms in terms of precision, recall and F-measures.**

Keywords*— *Local Binary Pattern, Midhinge, Abrupt transition, Adaptive threshold, Euclidean distance, Shot change detection.

I. Introduction

The surge in multimedia and network technology has driven to rapid growth in the amount of video data that is available to public now-a-days. This demands for efficient and effective tools for summarizing and retrieval of videos. Efficient Video Summarization and Retrieval based on content of the video requires shot change detection as a pilot step. Shots are uninterrupted frame sequences during camera breaks. Shots are identified by detecting shot boundaries or shot transitions (abrupt and gradual), where more than one shot are joined to form a scene. Abrupt transitions are immediate transitions between successive shots, whereas gradual transitions are slow and occurs over multiple frames [1]. Different video effects used for gradual transitions are fade-in, wipe, fade-out, dissolve etc. Earlier, varied techniques have been used to detect shot transitions. A comprehensive review and challenges of shot change detection is found in [1,2,3]. In literature, some of the SBD(Shot Boundary Detection) works has been emphasized based on edges [4,5], color histograms [6,7], motion features [7,8], pixels [9,10], similarity analysis [11], SIFT features [12] etc. Manjunath et al. [13] pro-posed a non-parametric method to detect shot boundaries using eigen gap approach. Features from orthogonal transform moments has been derived to detect hard cut in videos by Sadiq et al. [14]. Guru et al. [15] proposed a work for shot change detection based on split and merge framework using fisher linear discriminant criterion. Hui et al. [16] developed a fuzzy based technique to unify hybrid features to detect hard and gradual transition in a video. Zhang et al. [17] applied block-wise principal component analysis to establish shot eigen spaces from video segments to detect cut and gradual transition. Some of the drawbacks of the existing system states that edge based methods consumes more computational time [18] and color based methods in video segmentation gives motion induced false alarm and illumination variation [19]. In image processing applications like medical imaging [20], remote sensing, face identification and content-based image retrieval, texture feature plays a vital role. Prominent texture feature that is invariant to image rotation and illumination changes is Local Binary Pattern (LBP) [21]. Ojala et al. [22] proposed LBP method for texture classification. Further, to enhance the robustness and discriminative capability, different versions of LBP viz. Completed LBP [23], Mean LBP [24], Improved LBP [25], Mid-range LBP [26], Dominant LBP [27], Pyramid LBP [28] have been developed. Even though noticeable results on texture analysis have been attained by LBP and its variants, they often classify many patterns into a same class and are sensitive to noise. The proposed method overcomes this problem to certain extent using Midhinge LBP (MHLBP) which is an extension of LBP and it facilitate as the basic need in some of the video retrieval applications such as museum management, e-learning, remote sensing, architectural and engineering design, weather forecasting, geographic information systems etc.

The paper is organized as follows: section 2 provides brief review of MLBP and MRLBP. Section 3 discusses MHLBP feature extraction, video representation and shot boundary detection. Section 4 details experimental results and comparative analysis and section 5 provides the conclusion..

II. Related Works

The original LBP serves as a local descriptor which compares the central pixel value with its neighbouring pixel intensity values. Binary values 0/1 is designated by considering neighbouring pixel values exceeding (or equal to) centre pixel [29]. The original LBP operator has two significant drawbacks i.e., it's sensitivity towards noise and at times, it characterise various structural patterns with same binary code, which reduces its discriminative capability [30] as illustrated in Fig. 1. To improve the discriminative capability of basic LBP, several variants of LBP have been designed which includes Mean LBP (MLBP) and Midrange LBP (MRLBP). In this section MLBP and MRLBP methods are reviewed.

978-1-7281-4994-3/20/$31.00 ©2020 IEEE

A. Brief review of MLBP

Bai et al. [24] proposed MLBP method by analysing all the pixels in the block with the mean intensity value of 3*3 neighbourhood pixels except the centre pixel. Average gray value for each block is computed as in (1)

$$ALG = \sum_{i=0}^{8} G_i + G/9 \tag{1}$$

Where G denotes the gray value of the centre pixel and G_i (i=0,…,8) represents the gray values of the neighbourhood pixels. ALG stand for mean gray value and is used as a threshold instead of centre pixel, which is robust to noise. Comparison of mean threshold with neighbouring pixels is performed to obtain MLBP value as in (2):

$$MLBP_{(P\,R)} = \sum_{p=0}^{p-1} S(G_p - ALG_c)2^p \tag{2}$$

$$f(x) = \begin{Bmatrix} 1 & x \geq 0 \\ 0 & else \end{Bmatrix} \tag{3}$$

Where P is the number of pixels in the neighbourhood, R stands for the radius, G_P represents the neighbouring pixel gray values and ALG_C is mean gray value of 3*3 neighbourhood values. Fig. 2 illustrates an example of obtaining an MLBP from 3*3 pattern.

B. Brief Review of MRLBP

Rashmi and Nagendraswamy proposed MRLBP [26] method where the pixels in the block are compared with the midrange intensity value of 3*3 neighbourhood pixels except the centre pixel. Midrange threshold is obtained by evaluating mean of minimum and maximum values of 3*3 neighborhood pixels as shown in (4).

$$MRLG = \left(G_{min} + G_{max}\right)/2 \tag{4}$$

Here, G_{min} and G_{max} represents minimum and maximum values of 3*3 neighbour-hood. Comparison of midrange threshold with neighbouring pixels is performed to obtain MRLBP value as follows:

$$MRLBP_{(P\,R)} = \sum_{p=0}^{p-1} S(G_p - MRLG_c)2^p \tag{1}$$

$$f(x) = \begin{Bmatrix} 1 & x \geq 0 \\ 0 & else \end{Bmatrix} \tag{2}$$

Where P confines the number of pixels in the neighbourhood, R stands for the radius, G_P represents the neighbourhood values and $MRLG_C$ is mean gray value of 3*3 neighbourhood. Fig. 3 illustrates an example of obtaining an MRLBP from 3*3 pattern.

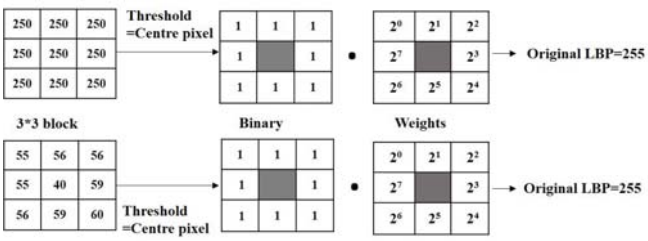


Fig. 1. Illustration of Original LBP operator

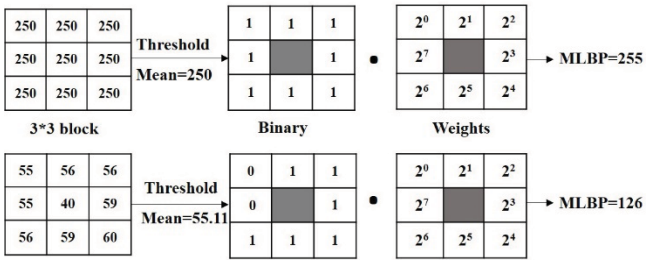


Fig. 2. Illustration of Mean LBP operator

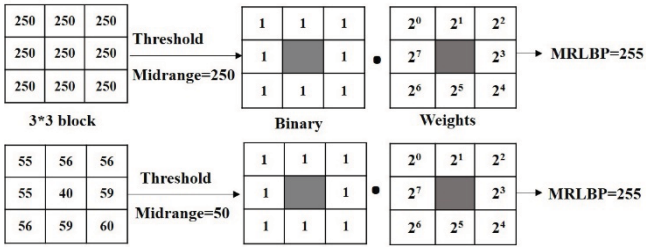


Fig. 3. Illustration of Midrange LBP operator

III. PROPOSED METHOD

The planned method detects abrupt shot changes in a video. Initially, histogram feature is constructed by extracting texture feature for each frame utilising MHLBP operator. The Euclidean Distance value is obtained between the histogram features of adjacent frames in a video. These values aids in determining adaptive threshold for detection of shot changes. Further, MHLBP texture description and video depiction scheme to address the detection of shot boundary are detailed in subsections. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Feature Extraction using Midhinge LBP (MHLBP)

The flexibility of LBP method makes it easily modifiable to address different types of problems. There are various enhancement of LBP that aims to enrich its robust-ness and discriminative capability. MLBP and MRLBP are one of those variants. Midhinge LBP has been proposed deriving motivation from MLBP and MRLBP. Discrimination capability of MHLBP is better than original LBP and its versions (mean LBP and midrange LBP). Unlike mean and midrange, the midhinge is relatively robust measure to estimate central tendency [31]. The average of first quartile and third quartile of 3*3 mask intensity values is considered as Midhinge threshold at each pixel position as follows:

$$MHLG = \left(Q1 + Q3\right)/2 \tag{3}$$

Where Q1 and Q3 are the first and third quartile gray intensity values of 3*3 neighbourhood. MHLG exhibits the midhinge gray value. A process in accordance to MRLBP is enforced using MHLG as the threshold in place of midrange gray value as formulated below:

$$MHLBP_{(P\,R)} = \sum_{p=0}^{p-1} S(G_p - MHLG_c)2^p \tag{8}$$

Here, p represents the pixels count in the neighbourhood, R depicts the radius, G_p gives gray values of neighbourhood and $MHLG_c$ is midhinge gray value of 3*3 mask.

The f(x) function is defined in the following equation:

$$f(x)=\begin{Bmatrix}1 & x\geq 0\\ 0 & else\end{Bmatrix} \tag{4}$$

The discriminative efficiency of MHLBP is illustrated in Fig. 4 which provides distinct MHLBP codes for 3*3 pattern when compared to Original LBP, Mean LBP and Midrange LBP as shown in Figs. 1, 2 and 3 respectively.

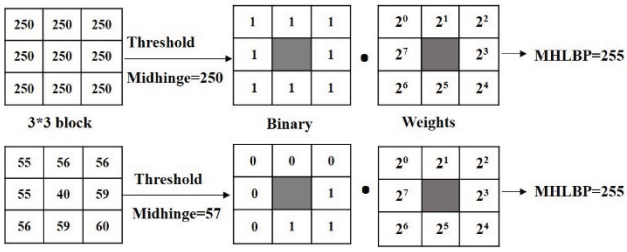


Fig. 4. Illustration of Midhinge LBP operator

B. Video Representation

As described in section A, the MHLBP operator provides codes for the overlying blocks of 3*3 matrix

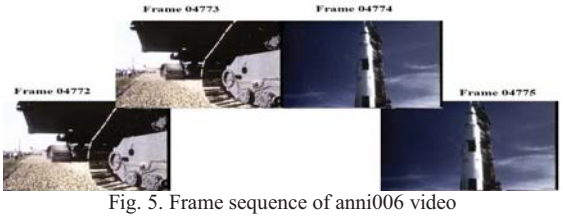


Fig. 5. Frame sequence of anni006 video

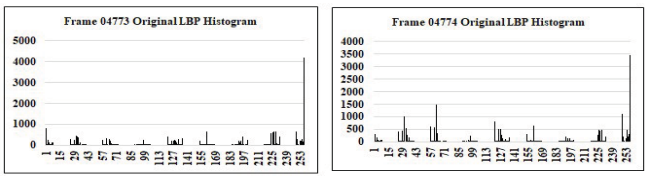


Fig. 6. Original LBP Histogram distribution

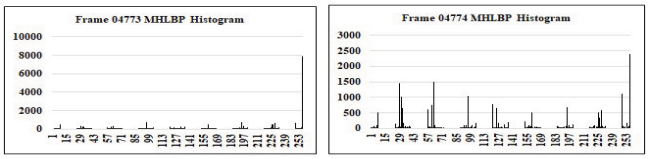


Fig. 7. MHLBP Histogram distribution

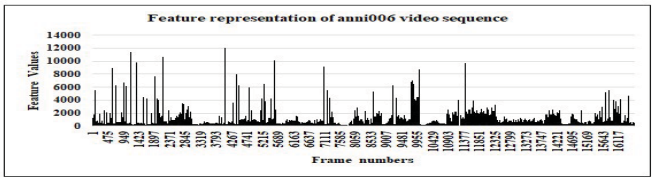


Fig. 8. Feature distribution of anni006 video

C. Shot Change Detection

Statistic-based metrics computed at the block level gives the best result for abrupt cut detection [32].In this approach, frame difference values are calculated by utilizing Euclidean distance measure for the histogram feature of the adjacent frames. Euclidean distance is computed as shown in equation 10.

$$D_{(p,q)}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \tag{10}$$

Where {Pi} and {qi} represents two histogram values of adjacent frames of a video, {i=1,2,….,n} and D gives the distance measure. Threshold is computed by considering the standard deviation (σ) and mean (μ) of the distance Di values for the entire frame sequence of a video as formulated:

$$Thd=(\mu+\sigma)*\alpha \tag{11}$$

Where α is a constant value. Abrupt cut is identified, if the dissimilarity between the frames exceeds the adaptive threshold value.

IV. Results

The planned method is analysed by performing experiments on few videos of TRECVID 2001 dataset. TRECVID is a part of TREC series which is a standard dataset for performance evaluation of various shot change detection algorithms [33]. These videos contains properties like cut, gradual transition, camera/object motion etc. The description of the dataset used in our experiment is described in Table 1 [34].

TABLE 1. Details of TRECVID 2001 dataset

video	Description	Number of Frames	Abrupt Transitions
anni005	"NASA_25th_Anniversary-Show_Segment_5"	11,363	38
anni006	"NASA_25th-Anniversary-Show_Segment_6"	12,306	41
anni009	"NASA_25th-Anniversary-Show_Segment_9"	16,587	38
NAD53	"A&S_Reports_Tape_#4_-_Report_#260"	26,115	83

The results obtained are analyzed with the available ground-truth to identify the shot transition as correctly identified shot cut, missed shot cut or falsely identified shot cut. The evaluation of the proposed method is computed using Precision, Recall and F-measure.

$$Precision=\frac{CI}{CI+FI} \tag{12}$$

$$\mathrm{Re}call=\frac{CI}{CI+MI} \tag{13}$$

$$F-Measure=\frac{2*\mathrm{Pr}ecision*\mathrm{Re}call}{\mathrm{Pr}ecision+\mathrm{Re}call} \tag{14}$$

A. Comparative analysis with LBP and Non-LBP methods

On the selected dataset, comparative study is performed with Original LBP, MLBP and MRLBP approaches and is represented in Table 2. The presented method provides significant results in terms of discriminative efficiency in comparison with original LBP and its variants. The performance is measured via recall, precision and F-measure. Also, comparative study is performed with few of the existing algorithms as depicted in Table 3. The presented method surpass some of the well-known approaches with the average performance parameter Recall score of 88%, Precision score of 90% and F-measure score of 89%. From the outcome of the experimental analysis, MHLBP descriptor has improved the ability of discrimination of original LBP and has better efficiency in comparison with

existing algorithms. Comparative analysis has been depicted in Fig. 9.

TABLE 1.COMPARATIVE ANALYSIS OF PROPOSED METHOD WITH OTHER METHODS

Video sequence	No. of shots	Original LBP [29]			Mean LBP [24]			Midrange LBP [26]			Proposed Method		
		R	P	Fm	R	P	Fm	R	P	Fm	R	P	Fm
"anni005"	38	0.76	0.94	0.84	0.79	0.86	0.82	0.76	0.88	0.82	0.86	0.89	0.87
"anni006"	41	0.66	0.77	0.71	0.73	0.77	0.75	0.83	0.83	0.83	0.87	0.90	0.88
"anni009"	38	0.74	0.88	0.80	0.74	0.90	0.81	0.79	0.86	0.86	0.89	0.91	0.89
"NAD53"	83	0.77	0.77	0.77	0.81	0.81	0.81	0.87	0.87	0.87	0.91	0.92	0.91
Average		0.73	0.84	0.78	0.77	0.84	0.80	0.81	0.86	0.84	0.88	0.90	0.89

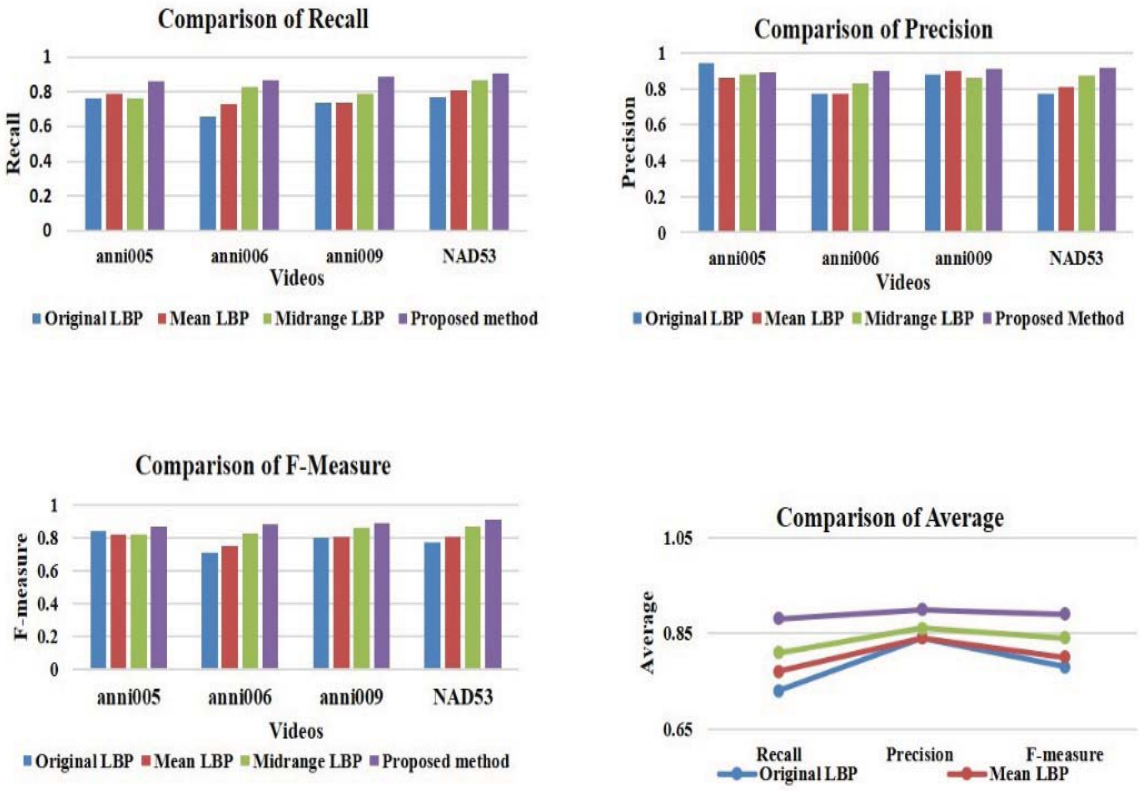


Fig. 1. Graph for Recall, Precision, F-measure and Average

TABLE 2.COMPARATIVEANALYSIS OF PROPOSED METHOD WITH NON-LBP BASED METHODS TO DETECT SHOT BOUNDARIES

Video Sequence	Number of shots	Color Histograms [35]			Motion Vector Likelihoods [36]			Proposed Method		
		R	P	Fm	R	P	Fm	R	P	Fm
anni005	38	0.83	0.64	0.72	0.53	0.46	0.49	0.86	0.89	0.87
anni006	41	0.70	0.78	0.74	0.56	0.57	0.56	0.87	0.90	0.88
anni009	38	0.71	0.84	0.77	0.64	0.59	0.61	0.89	0.91	0.89
NAD53	83	0.62	0.69	0.65	0.73	0.46	0.56	0.91	0.92	0.91
Average		0.72	0.74	0.72	0.61	0.52	0.56	0.88	0.90	0.89

V. CONCLUSION

In this work, abrupt transition of videos are identified by extracting texture features from video frames using midhinge LBP. The performance of the LBP feature is improved by extracting MHLBP histogram for every frame of a video. MHLBP has enhanced the discriminative ability compared to original LBP and its versions (MLBP and MRLBP). The proposed method has been tested with some of the existing methods on TRECVID2001 video dataset and the results obtained are acceptable in terms of efficiency and accuracy. The outcome of the proposed method provided 89% of F-measure whereas using the LBP variants viz. original LBP, Mean LBP, Midrange LBP we obtained 78%, 80% and 84% of F-measure respectively. Further research is to be focused on the detection of gradual transitions in the video dataset.

REFERENCES

1. Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., Zhang, B.: A formal study of shot boundary detection. In: IEEE transactions on circuits and systems for video technology, vol.17 (2), pp.68-186. (2007).
2. Abdulhussain, S. H., Ramli, A. R., Saripan, M. I., Mahmmod, B. M., Al-Haddad, S. A. R., Jassim, W. A.: Methods and challenges in shot boundary detection: a review. Entropy, 20(4), 214. (2018).
3. Hanjalic, A.: Shot-boundary detection: unraveled and resolved?. In: IEEE transactions on circuits and systems for video technology, vol. 12(2), pp. 90-105. (2002).
4. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying production effects. In: Multimedia systems, vol. 7(2), pp. 119-128. (1999).
5. Adjeroh, D., Lee, M. C., Banda, N., Kandaswamy, U.: Adaptive edge-oriented shot boundary detection. In: EURASIP Journal on Image and Video Processing, (2009).
6. Swain, M. J., Ballard, D. H.: Color indexing. In: International journal of computer vision, vol. 7(1), pp. 11-32. (1991).
7. Adjeroh, D. A., Lee, M. C.: Robust and efficient transform domain video sequence analysis: An approach from the generalized color ratio model. In: Journal of Visual Communication and Image Representation, vol. 8(2), pp. 182-207. (1997).
8. Dhawale, C. A., Jain, S.: Motion compensated video shot detection using multiple feature experts. IN: ICGST Int. J. Graph. Vis. Image Process. GVIP, Vol. 8, pp. 1-11. (2008).
9. Bouthemy, P., Gelgon, M., Ganansia, F.: A unified approach to shot change detection and camera motion characterization. In: IEEE transactions on circuits and systems for video technology, vol. 9(7), pp. 1030-1044. (1999).
10. Zhang, H., Kankanhalli, A. Smoliar, S.W.: Automatic partitioning of full-motion video. In: Multimedia systems, vol. 1(1), pp.10-28. (1993.)
11. Mohanta, P. P., Saha, S. K., Chanda, B.: A model-based shot boundary detection technique using frame transition parameters. In: IEEE Transactions on multimedia, vol. 14(1), pp. 223-233. (2011).
12. Chang, Y., Lee, D. J., Hong, Y., Archibald, J.: Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor. In: EURASIP Journal on Image and Video Processing, vol. 2008(1), 860743. (2007).
13. Manjunath, S., Guru, D. S., Suraj, M. G., Harish, B. S.: A non parametric shot boundary detection: an eigen gap based approach. In: Proceedings of the Fourth Annual ACM Bangalore Conference, pp. 14. (2011).
14. Abdulhussain, S. H., Ramli, A. R., Mahmmod, B. M., Saripan, M. I., Al-Haddad, S. A. R., Jassim, W. A.: Shot boundary detection based on orthogonal polynomial. In:Multimedia Tools and Applications, pp.1-22. (2019).
15. Guru, D. S., Suhil, M., Lolika, P.: A novel approach for shot boundary detection in videos. In: Multimedia Processing, Communication and Computing Applications, pp. 209-220. Springer, (2013).
16. Fang, H., Jiang, J., Feng, Y.: A fuzzy logic approach for detection of video shot boundaries. In: Pattern Recognition, vol. 39(11), PP. 2092-2100. (2006).
17. Zhang, D., Lei, W., Zhang, W., Chen, X.: Shot boundary detection based on block-wise principal component analysis. In: Journal of Electronic Imaging, vol. 28(2), pp-023029. (2019).
18. Domnic, S.: Walsh–Hadamard transform kernel-based feature vector for shot boundary detection. In: IEEE Transactions on Image Processing, vol. 23(12), pp. 5187-5197. (2014).
19. Dadashi, R., Kanan, H. R.: AVCD-FRA: A novel solution to automatic video cut detection using fuzzy-rule-based approach. In: Computer Vision and Image Understanding, vol. 117(7), pp. 807-817. (2013).
20. Chakraborty, C.: Computational approach for chronic wound tissue characterization. Informatics in Medicine Unlocked. p.100162. (2019).
21. Mäenpää, T., Pietikäinen, M.: Texture analysis with local binary patterns. In: Handbook of pattern recognition and computer vision, pp. 197-216. (2005).
22. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. In: Pattern recognition, vol. 29(1), pp. 51-59. (1996).
23. Zhao, Y., Jia, W., Hu, R. X., Min, H.: Completed robust local binary pattern for texture classification. In: Neurocomputing, vol. 106, pp. 68-76. (2013).
24. Bai, G., Zhu, Y., Ding, Z.: A hierarchical face recognition method based on local binary pattern. In: 2008 congress on image and signal processing, vol. 2, pp. 610-614. IEEE, (2008).
25. Jin, H., Liu, Q., Lu, H., Tong, X.: Face detection using improved LBP under Bayesian framework. In: Third International Conference on Image and Graphics, pp. 306-309. IEEE, (2004).
26. Rashmi, B. S., Nagendraswamy, H. S.: Video shot boundary detection using midrange local binary pattern. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 201-206. IEEE, (2016).
27. Liao, S., Law, M. W., Chung, A. C.: Dominant local binary patterns for texture classification. In: IEEE transactions on image processing, vol. 18(5), pp. 1107-1118. (2009).
28. Qian, X., Hua, X. S., Chen, P., Ke, L.: PLBP: An effective local binary patterns texture descriptor with pyramid representation. In: Pattern Recognition, vol. 44(10-11), pp. 2502-2515. (2011).
29. Pietikäinen, M., Zhao, G.: Two decades of local binary patterns: A survey. In: Advances in independent component analysis and learning machines, pp. 175-210. Academic Press (2015).
30. Kaya, Y., Ertuğrul, Ö., Tekin, R.: Two novel local binary pattern descriptors for texture analysis, In: Applied Soft Computing, vol. 34 (2015), pp. 728-735. (2015).
31. Doane, D. P., Seward, L. W.: Applied statistics in business and economics. McGraw-Hill/Irwin, New York, NY (2011).
32. Satapathy, S. C., Govardhan, A., Raju, K. S., Mandal, J. K.: Emerging ICT for Bridging the Future. In: Proceedings of the 49th Annual Convention of the Computer Society of India (CSI), Vol. 1. Springer, (2014).
33. https://www-nlpir.nist.gov/projects/tv2015/index.html
34. https://www-nlpir.nist.gov/projects/trecvid/collection.html
35. Adjeroh, D., Lee, M. C., Banda, N., Kandaswamy, U.: Adaptive edge-oriented shot boundary detection. EURASIP Journal on Image and Video Processing, 2009(1), 859371 (2009).
36. Li, W. K., Lai, S. H.: Integrated video shot segmentation algorithm. In: Storage and Retrieval for Media Databases International Society for Optics and Photonics,vol. 5021, pp. 264-271. (2003).

ARTICLE IN PRESS

Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Shot based keyframe extraction using edge-LBP approach

H.M. Nandini [a,b,*], H.K. Chethan [a], B.S. Rashmi [b]

[a] Department of Computer Science and Engineering, Maharaja Research Foundation, Maharaja Institute of Technology, Mysuru, India
[b] Department of Information Technology, Karnataka State Open University, Mysuru, India

ARTICLE INFO

Article history:
Received 2 September 2020
Revised 28 October 2020
Accepted 31 October 2020
Available online xxxx

Keywords:
Sobel operator
Shot boundary detection
Gradient function
Keyframe extraction
Z-score

ABSTRACT

Advancement in technology has led to tremendous increase in the online video content that requires efficient and effective content based video analysis approaches. In this regard, efficient approach for abrupt Shot Boundary Detection (SBD) and keyframe extraction has been presented. The proposed method detects abrupt shots by extracting binarized edge information from frames for texture characterisation using Local Binary Pattern (LBP) method. Further, Euclidean distance has been applied on the histogram features constructed and an adaptive threshold is used to detect abrupt shots. During keyframe extraction phase, magnitude gradient using Sobel operator has been extracted from each frame of the segmented shot. Subsequently, magnitude values are transformed into Z-score which describes the position of each pixel in terms of its distance from the mean, when measured in standard deviation units of every frame. Finally, Co-efficient of variation is computed for each frame and the frame possessing the highest value is selected as a keyframe from every shot. Experiments were conducted on TRECVID 2001 dataset to analyze and validate the proposed approach. Experimental result manifest that the proposed SBD and keyframe extraction method outperforms some of the state-of-the-art algorithms with average F1-score of 98.15% and average fidelity measure of 90% respectively.

1. Introduction

Nowadays, the rapid development of multimedia technology has led to increase in the number of video data available on the web. This demands the need for new approaches to facilitate video summarization, indexing and retrieval. Video shots are often treated as the basic element for video analysis (Duan et al., 2013). A shot is a consecutive frame sequence that are taken from single camera break (Thakre et al., 2016). Basically, shot boundaries are categorized into abrupt and gradual transitions. Gradual transitions includes dissolve, fade-in, fade-out and wipes (Yuan et al., 2007). The entire video shot can be represented using a single frame, called keyframe (Hannane et al., 2016).

Video summarization depicts compact representation of a video sequence (Besiris et al., 2008). It is represented as a sequence of still images (keyframes) or moving images (video skims) (Truong and Venkatesh, 2007). Advantages of video skim over static summary is the ability to include audio and motion elements that enhances the information to be conveyed by the summary (Jadhav and Jadhav, 2015). In static summary, keyframes are not restricted with time and sequence issue. So, it provides more flexibility in terms of arrangement for the purpose of browsing and navigation. Therefore, in recent years researchers are focusing more on efficient approaches to develop static summaries.

In this paper, we have focused on generating static video summary by extracting keyframes from the segmented shots of the video. There are numerous ways to extract features based on texture, shapes, edges etc. Texture based extraction plays a vital role in image processing applications like medical image analysis, remote sensing, face identification and content-based image retrieval. Most generally used texture descriptor is Local Binary Pattern (LBP) (Pietikainen, 2005). Further, to improve the robustness and discriminative capability, different versions of LBP viz. Mean LBP (Bai et al., 2008), Improved LBP (Jin et al., 2004), Mid-range (Rashmi, 2016), Dominant LBP (Liao et al., 2009), Pyramid LBP (Qian et al., 2011) and others (Sliti et al., 2018) (Khaleefah et al., 2019) have been developed. Despite the fact that remarkable results on texture analysis have been gained by LBP and its variants, they often classify many patterns into a same class and are

* Corresponding author.
E-mail addresses: nandiniksou@gmail.com (H.M. Nandini), hkchethan@gmail.com (H.K. Chethan), rashmibsrsh@compsci.uni-mysore.ac.in (B.S. Rashmi).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2020.10.031
1319-1578/

Please cite this article as: H.M. Nandini, H.K. Chethan and B.S. Rashmi, Shot based keyframe extraction using edge-LBP approach, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.10.031

sensitive to noise. To detect the occurrence of shot boundary accurately in the video sequence, we must select effective and efficient features from the frames of the video which are robust to misleading factors such as illuminations or lighting effects, zooming etc. Edges are largely invariant under local illumination changes and are much less affected by possible motion in the video (Adjeroh et al., 2009). The recent literature related to LBP shows that the accuracy of the original LBP method can be improved using edge information (Abdesselam, 2013). This has motivated us to carry out this research work.

The proposed method detects abrupt transitions based on LBP texture histograms that are constructed using binarized edge information of the frames. This has enhanced the accuracy of the original LBP method. The distance between each adjacent histogram of frames is computed using Euclidean Distance and adaptive threshold is applied to detect the shot boundaries. In the subsequent step, keyframe extraction is performed by applying Sobel gradient function by estimating its magnitude on each frame of a shot. Then, Z-score is applied on the magnitude values and co-efficient of variation is computed to extract the keyframes. Finally, video summary is devised by combining the keyframes obtained.

The rest of the paper is organized as follows: Section 2 presents the brief description of some related works. Section 3 gives the detailed description of the proposed methodology for SBD and keyframe extraction. Section 4 reports the experimental analysis and results and Section 5 concludes the paper.

2. Literature review

In this section, a brief overview of literature on SBD and keyframe extraction approaches by various researchers is presented.

2.1. Shot boundary detection

In literature, various approaches for SBD have been explored (Yuan et al., 2007; Abdulhussain et al., 2018). Most of the approaches are based on histogram, edge, pixel etc. Huo et al. (2016) have proposed poisson model based inter-frame difference histogram method to detect cut and gradual transitions in video. A novel algorithm to detect abrupt transitions via dual stage approach has been presented (Singh et al., 2020). In the first stage adaptive wiener filter is applied and Local Binary Pattern histogram Fourier is extracted to reduce the effect of illumination. In the second stage Canny edge difference is used to remove the motion and illumination effects. Orthogonal transform moments (Abdulhussain et al., 2019) have been used to extract feature from the frame to detect hard cuts in the video. Rashmi (2016) have introduced Midrange Local Binary Pattern (MRLBP) as a texture descriptor to detect the abrupt cuts in the video. Scale invariant feature transform have incorporated to the RGB color space in order to identify the abrupt and gradual transitions in the video (El khattabi et al., 2017). A novel approach for identifying shot cuts in the video based on split and merge frame work using Fischer Linear Discriminant Criterion has been presented (Guru et al., 2013). Hannane et al., (2016) have proposed an efficient method to detect abrupt and gradual transitions in video by extracting the SIFT-Point distribution histogram from the frames, which is a combination of both global and local features. Bitwise-XOR dissimilarity operation between the adjacent frames of the video has been utilised (Rashmi and Nagendraswamy, 2016) to detect abrupt cuts . Dadashi and Kanan (2013) have presented a new method to detect abrupt cuts in the video based on fuzzy rules. Motion based SBD (MSBD) has been presented (Kanagaraj and Priya, 2018) using curvlet features as it represents object motion in different magnitude and orientations.

In spite of the useful previous works in the literature for SBD, it is still a challenging issue to attain quality performance for all genre videos.

2.2. Keyframe extraction

Extraction of keyframes plays a vital role in video summarization, indexing and retrieval. Hence, lot of research has been carried out for keyframe extraction. As reported by Angadi and Naik (2014), there are mainly four different types of approaches for keyframe extraction: sampling based approach, object based approach, segment based approach and shot based approach. In sampling based technique, keyframes are selected randomly under sampling without considering content of the video. Main disadvantage of this technique is that, it may cause some necessary yet short video clips to have no representative frames. Object based approaches are efficient and semantic, but more prominence is given to the foreground and are convenient for certain applications only. The keyframes selected by segment based technique could efficiently represent the content of the video. Even so, segmentation is a complex process and it is hard to decide the number of segments. Shot based approaches are effective and one of the important method as they select keyframes from each shot.

Hannane et al. (2016) have proposed a technique to extract keyframes from the shots using entropy based singular value metric. Probabilistic entropy measure has been utilized (Rashmi and Nagendraswamy, 2018) to choose keyframe within fuzzified frames of a video shot. Pan et al. (2019) introduced keyframe extraction approach based on clustering. Keyframes are selected based on their energy rank derived from dissimilarity and representativeness of video frames. A novel video summarization approach based on Color co-occurrence matrices for SBD is performed using normalised sum of squared differences and middle frame of shot is selected as a keyframe (Mussel Cirne and Pedrini, 2018). Computational framework that makes use of one-class classifier as a novelty detector is used (Yong et al., 2013) to extract keyframes based on the semantic-contents of the frame. Barhoumi and Zagrouba (2013) have presented a method to extract keyframes from the shots based on object based event detection. A new method to extract keyframes based on logical image description using interest points, repeatability network and modularity has been designed (Gharbi et al., 2019). Kanagaraj and Priya (2018) have presented shot based keyframe extraction for feature extraction and selection for multimedia event classification. Here, Block Matching Algorithm (BMA) is performed to avoid similarity among keyframes.

Above discussed methods makes it clear that various techniques are available to extract keyframes for video summarization. Still, it is a big challenge to extract keyframes from the segmented video to construct an efficient video summary.

3. Proposed methodology

SBD and keyframe extraction are two important phases of the proposed methodology. The preliminary step is to convert the extracted video frames from color to gray before moving forward to afore said phases.

Initially, gray scale images are transformed into binarized edge images using different edge detection operators like Sobel, Canny and Roberts. Then, block based LBP is computed on the obtained binarized images with and without centre pixel and histogram is constructed to form a feature set. Further, Euclidean Distance is applied between the histogram features of adjacent frames. Adaptive threshold is computed on the obtained distance values to identify the shot boundaries. In the subsequent step, Sobel gradient

function is applied to each frame of the segmented shot and its magnitude is estimated. Z-score is applied on the obtained magnitude gradient image and co-efficient of variation is computed. Finally, frame with maximum co-efficient of variation is selected as keyframe from every shot.

The above phases are broadly classified into three steps namely: (i) Feature extraction and representation (ii) Abrupt shot detection and (iii) Keyframe extraction. The following subsection describes each step of the proposed framework. Fig. 1 depicts the general framework of the proposed methodology.

3.1. Feature extraction and representation

Features that describes the visual information plays a major role to identify the occurrence of shot boundary. The original LBP operator makes use of the centre pixel to threshold each image pixels of 3*3 neighborhood. Binary values 0/1 is designated based on the threshold. LBP code is generated for each 3*3 neighborhood by multiplying the binary values with corresponding weights and summing up the result (Ojala et al., 1996). A new feature extraction approach, Binarized Edge Local Binary Pattern (BELBP) has been proposed in this work. Initially, the gray scale images are transformed into binarized edge images using Sobel (Sobel and Feldman, 1968), Roberts (Roberts, 1963) and Canny (Canny, 1986) edge detection operators as shown in Fig. 2.

These edge descriptors are used during experimentation to investigate their performance by applying LBP method. The performance of the LBP based methods are improved by obtaining the feature vectors from binarized edge image pixels. Here, feature vectors are constructed by applying LBP technique with and without centre pixel of 3×3 neighborhood for the binarized edge images. Since binarized edge image is considered, threshold computation step is eliminated. LBP code computed lies in the range 0 to 511 considering centre pixel and 0–255 for not considering centre pixel. The BELBP is computed as formulated in the equation:

$$BELBP_{(P\ R)} = \sum_{p=0}^{p-1} S(G_p)2^p \tag{1}$$

Here, p represents the pixels count in the neighborhood, R depicts the radius, G_p gives binary values of 3*3 mask of binarized edge image. Thus, histogram is constructed using these LBP codes for each frame which aids in the detection of shot boundary. Fig. 3 shows the illustration of computing the LBP code for binarized Sobel edge image.

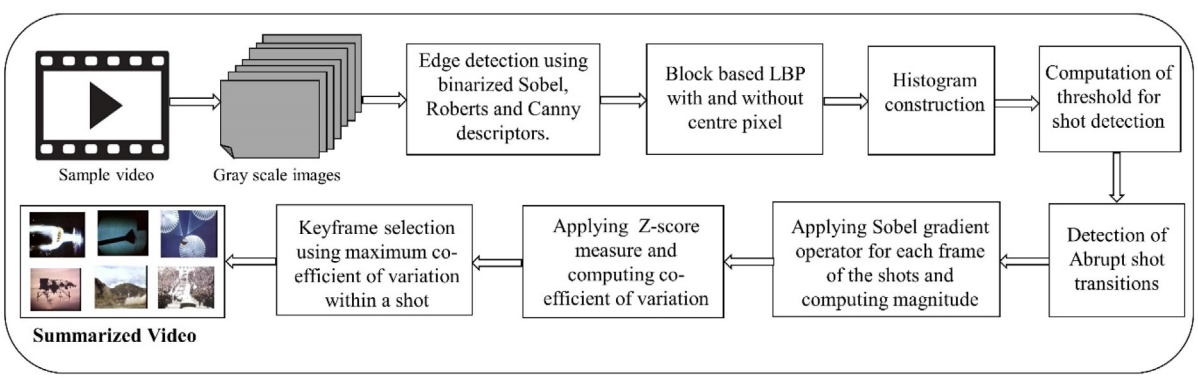


Fig. 1. Framework of the proposed methodology.

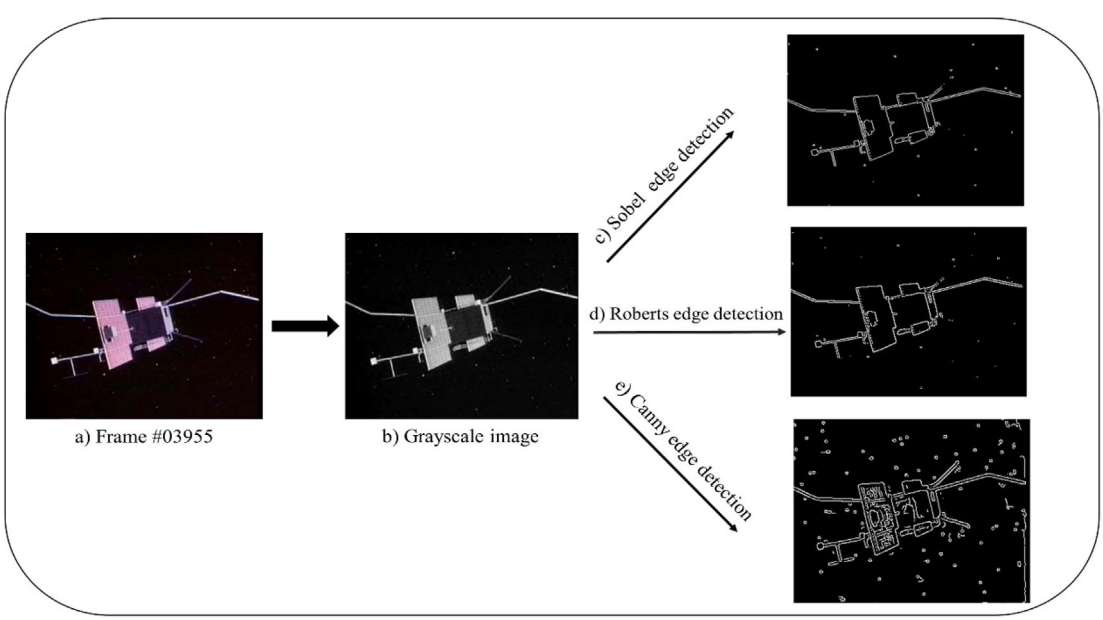


Fig. 2. Illustration of binarized edge images for frame #03955 of anni009 video.

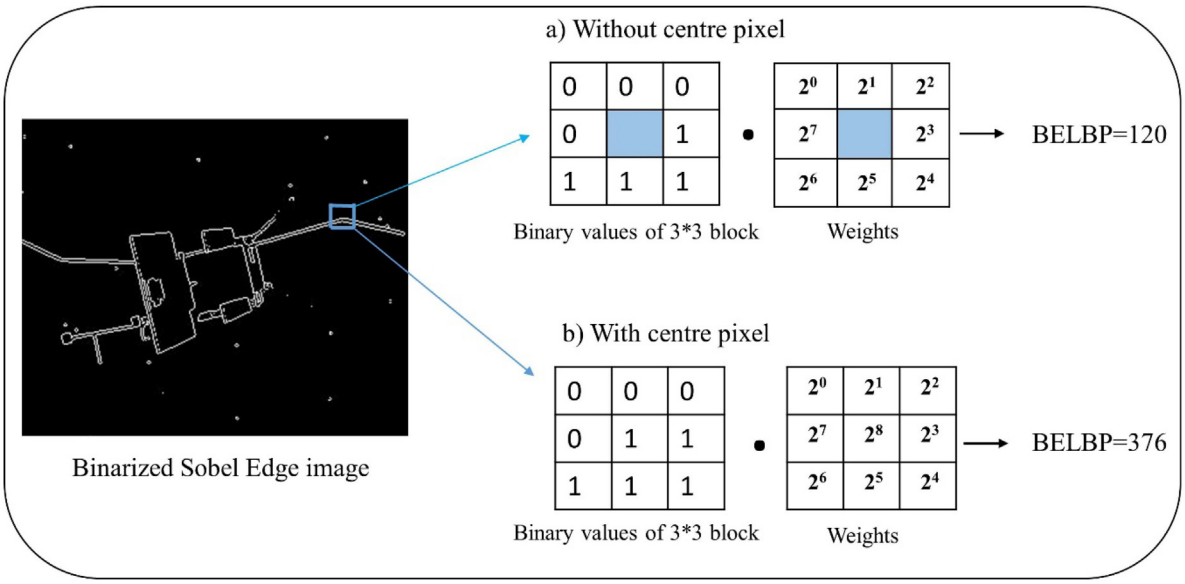


Fig. 3. Illustration of computing Sobel BELBP for frame #03955 of anni009 video.

3.2. Abrupt shot detection

After extracting the histogram features, the next step is to analyze the appropriate metrics to compare the adjacent frames. During abrupt shot detection, best results are obtained by using statistic based metrics computed at the block level (Ford et al., 2000). In the proposed method, we have utilized Euclidean Distance measure to compute the histogram frame difference values of the adjacent frames. Euclidean Distance computed is represented as in the equation.

$$DM_{(r,s)} = \sqrt{\sum_{i=1}^{n} (r_i - s_i)^2} \tag{2}$$

where DM represents the distance measure and $\{r_i\}$,$\{s_i\}$ are the two histogram values of adjacent frames of a video, where $\{i = 1,2,3.\ldots, n\}$.

On obtaining the distance values for the frame sequence of a video, adaptive threshold is computed. In this experiment, Standard Deviation (σ) and Mean (μ) of the distance values are considered to compute the threshold and is formulated in the following equation.

Where, A_{thd} provides the adaptive threshold and α is a constant value. Abrupt shot change is identified if the dissimilarity between the frames exceeds the adaptive threshold value.

3.3. Keyframe extraction

After detecting the abrupt shot boundary in the video, keyframe is extracted which efficiently reflects the salient content of the shot. Further, extracted keyframes are combined to form a video summary. In literature, there are various techniques to choose keyframes with the justifiable quality but they are computationally expensive (Furini et al., 2010). In this method, Sobel gradient operator is chosen to transform gray scale image to gradient magnitude image as it outperforms other edge detectors in terms of accuracy and computational efficiency (Abdesselam, 2013). In order to extract keyframes, we first apply gradient function using Sobel operator to each frame of the segmented shot. The operator consists of a pair of 3*3 kernels which are convolved with original image to calculate approximations of the derivatives, one in the x-direction (columns) and the other estimating the gradient in the y-direction (rows) as shown in the following equation.

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}$$

$$G_y = \begin{bmatrix} +1 & 2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{4}$$

Separate measurements are produced by the kernels for the gradient component in each direction, namely G_x and G_y. The magnitude of the gradient is then computed using below equation and Fig. 4 represents the magnitude of Sobel gradient image.

$$GM = \sqrt{G_x^2 + G_y^2} \tag{5}$$

Then, magnitude gradient values are transformed into Z-score that measures the position of each pixel of an image in terms of its distance from the mean, in units of standard deviation as formulated in the following equation:

$$ZGM = (GM_{ij} - \mu)/\sigma \tag{6}$$

where, $\mu = \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} GM_{ij}\right)$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(GM_{ij} - \mu)^2}{n}}$$

Further, co-efficient of variation is computed which aids in identifying the frames with rich visual content. The outcome of highest co-efficient of variation value indicates that the pixels in the frame/image are prominent in the edge region and the smallest co-efficient of variation indicates that pixel belongs to uniform region (Badshah et al., 2012). Based on this criteria, the frame representing the highest co-efficient of variation value is thus selected as keyframe within a video shot and is formulated as follows:

$$CV = \sigma/\mu \tag{7}$$

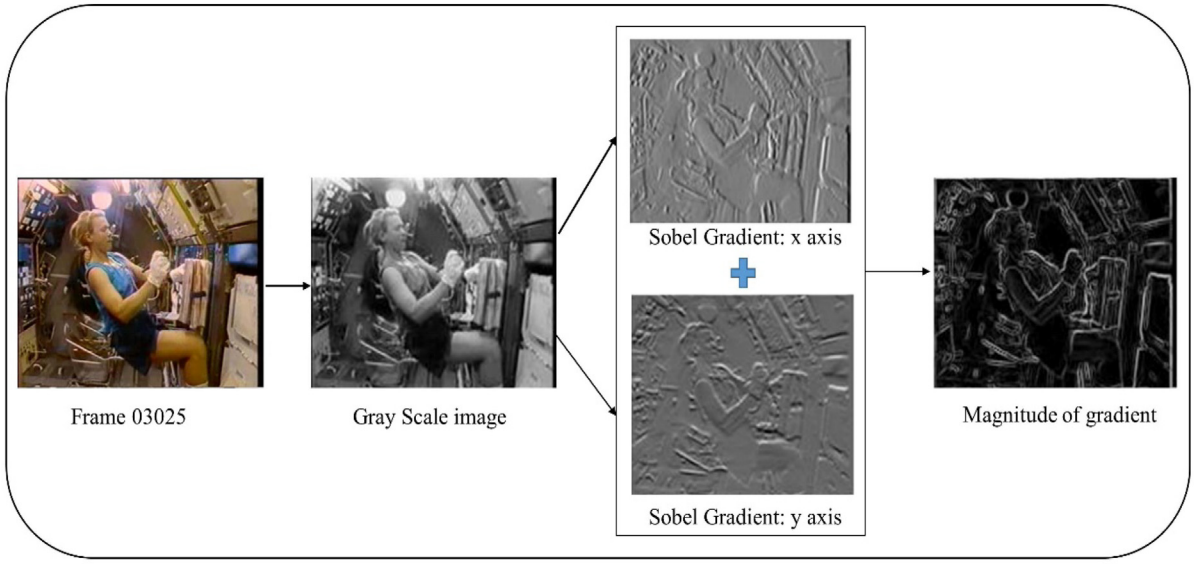


Fig. 4. Illustration of computing the magnitude of Sobel gradient image.

where, $\mu = \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} ZGM_{ij}\right)$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(ZGM_{ij}-\mu)^2}{n}}$$

4. Results and discussion

Dataset: In order to analyze the performance of the proposed method, video sequences from TRECVID 2001 dataset has been considered and description is given in Table 1. The dataset consists of abrupt cuts and different types of gradual transitions along with object motion, camera motion and illumination variation of various degree, which makes detection process very tedious. The frame resolution of TRECVID 2001 video sequences is 320 × 262. To analyze the proposed method, the shot transition information of these videos available with the ground truth is taken into consideration.

The proposed method is evaluated at two stages: i) Shot Boundary Detection and ii) Keyframe extraction. Extensive experiments were conducted to analyze the performance of the proposed technique using Matlab 2014a on Intel Core i5 processor, running at 2.70 GHz with 8 GB RAM.

Table 1
Description of TRECVID 2001 video dataset.

File Name	Video Title	No. of Frames	Abrupt transitions
anni005	"NASA_25th_Anniversary-Show_Segment_5"	11,363	38
anni006	"NASA_25th-Anniversary-Show_Segment_6"	12,306	41
anni009	"NASA_25th-Anniversary-Show_Segment_9"	16,587	38
anni010	"NASA_25th-Anniversary-Show_Segment_10"	31,389	98
nad31	"Spaceworks - Episode 6"	52,405	187
nad33	"Spaceworks - Episode 8"	49,768	189
nad53	"A&S_Reports_Tape_#4_-_Report_#260"	26,115	83
nad57	"A&S_Reports_Tape_#4_-_Report_#264"	12,781	44

4.1. Performance evaluation for abrupt SBD

During SBD, the obtained results are compared with ground truth, to classify the shot detection as correct, false or missed cuts and following measures were used to evaluate the performance:

$$Recall = \frac{C_d}{C_d + M_d} \tag{8}$$

$$Precision = \frac{C_d}{C_d + F_d} \tag{9}$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{10}$$

where C_d is the number of correctly identified shot cuts, M_d is the number of missed shot cuts and F_d represents false cuts. An outstanding shot transition detection algorithm must acquire high recall and high precision values.

Experiments have been conducted to evaluate the proposed method on standard video dataset as represented in Table 1 by considering six different cases. In the proposed method, systematic experimentations have been carried out on different combinations as shown in Fig. 5.

The experimental analysis are classified based on Sobel, Roberts and Canny edge images as follows: 1) Results of LBP with centre pixel 2) Results of LBP without centre pixel.

LBP with centre pixel: LBP has been applied on binarized Sobel, Roberts and Canny edge images considering the centre pixel of 3*3 neighborhood and histogram is constructed. The adaptive threshold computed are applied on histogram feature vectors of adjacent frames. The results analyzed in terms of recall, precision and F1-score for Sobel, Roberts and Canny images are detailed in Table 2.

Obtained results exhibits that, the performance of Sobel BELBP with centre pixel is better than other two methods in terms of recall, precision and F1-score.

LBP without centre pixel: Similarly, LBP has been applied to the above mentioned binarized edge images without centre pixel of 3*3 neighborhood and histogram is constructed. The adaptive threshold computed are applied on histogram feature vectors of adjacent frames. The results analyzed in terms of recall, precision and F1-score for Sobel, Roberts and Canny images are detailed in Table 3.

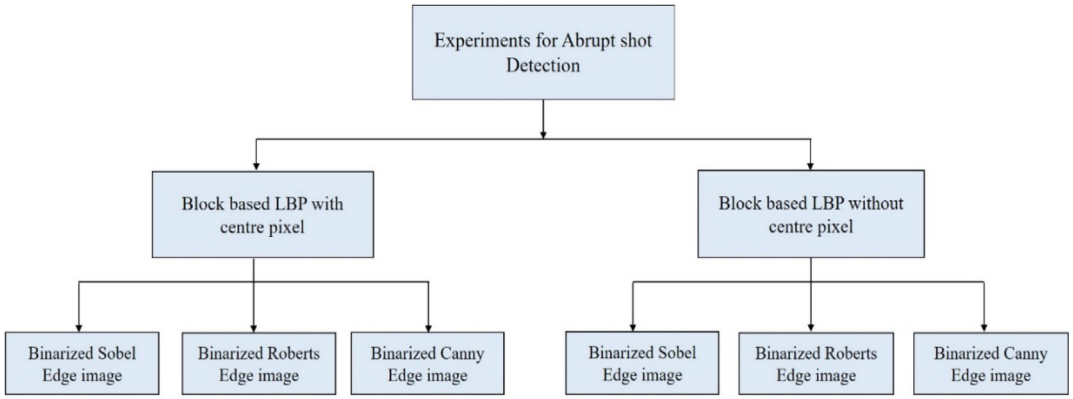


Fig. 5. Experimental setup for the proposed abrupt shot change method.

Table 2
Results of abrupt SBD using LBP without centre pixel for different edge operators.

Video Sequence	Sobel Edge image			Roberts Edge image			Canny Edge image		
	R	P	F	R	P	F	R	P	F
anni005	100	95	97.44	100	92.68	96.20	97.36	90.24	93.66
anni006	97.56	93.02	95.23	95.12	90.69	92.85	92.68	88.37	90.47
anni009	100	97.43	98.69	94.73	92.30	93.49	92.10	89.74	90.90
nad53	98.8	96.47	97.62	96.38	94.11	95.23	93.97	91.76	92.85
nad57	100	97.77	98.87	93.18	89.13	91.11	90.90	86.95	88.88
Average	99.27	95.93	97.57	95.88	91.78	93.77	93.40	89.41	91.35

Table 3
Results of abrupt SBD using LBP with centre pixel for different edge operators.

Video Sequence	Sobel Edge image			Roberts Edge image			Canny Edge image		
	R	P	F	R	P	F	R	P	F
anni005	97.36	88.09	92.49	94.73	87.80	91.13	92.10	87.5	89.74
anni006	92.68	86.36	89.40	90.24	86.04	88.08	90.24	85.71	87.91
anni009	94.73	90.88	92.76	92.10	87.50	89.99	89.47	87.17	88.30
nad53	96.38	90.90	93.55	93.97	90.69	92.30	92.77	89.53	91.12
nad57	90.90	88.88	89.88	90.90	86.95	88.88	88.63	86.66	87.63
Average	94.41	89.02	91.61	92.38	87.79	90.07	90.64	87.31	88.94

Comparative analysis in Table 3 shows that Sobel BELBP without centre pixel outperforms the other two methods in terms of recall, precision and F1-score.

The analysis of results in Table 2 and Table 3 depicts that, Sobel BELBP without considering centre pixel performs better than all the experimental cases in terms of recall, precision and F1- score and is represented graphically in Fig. 6.

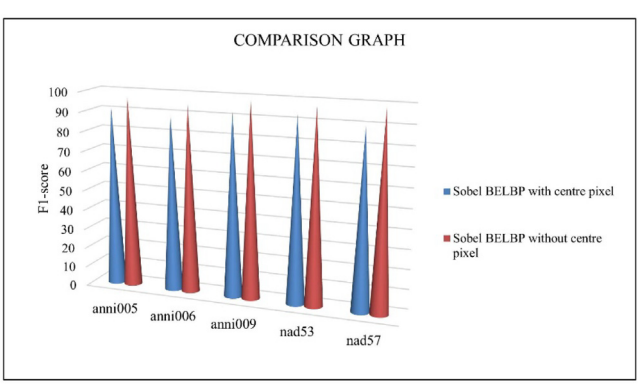


Fig. 6. Comparison graph for the best results obtained.

The CPU time taken is recorded for each method is tabulated in Table 4 to analyze the performance. According to the analysis, time taken by the Sobel BELBP without centre pixel is less when compared to all other methods.

As a conclusion, it is observed that Sobel BELBP without centre pixel is considered as an efficient method in terms of accuracy and computational speed when compared to other methods. The best case discussed is compared with other state-of-the-art LBP based SBD methods as shown in Table 5. The proposed method outperforms original LBP (Ojala et al., 1996) and one of its variant MRLBP (Rashmi, 2016) in terms of discriminative efficiency with 97% F1 score and it is graphically represented in Fig. 6.

Also, comparative study is performed with some of the existing non-LBP state-of-the-art SBD algorithms. The results in Table 6 shows the improved efficiency of the proposed method with non-LBP state-of-the art methods. The presented method surpass SBD method using dual stage SBD technique (Chakraborty and Thounaojam, 2020), soft computing techniques (Rashmi and Nagendraswamy, 2018) and SIFT-point distribution histogram method (Hannane et al., 2016) with 98.15% F1 score and the performance is represented graphically in Fig. 7.

However, the proposed algorithm is computationally expensive than global histogram approaches (Fig. 8). Detection of shot boundaries in proposed method is threshold dependent and is

Table 4
CPU time taken (milliseconds) per frame for feature extraction of different methods.

Video Sequence	LBP with centre pixel			LBP without centre pixel		
	Sobel edge	Roberts edge	Canny edge	Sobel edge	Roberts edge	Canny edge
anni005	0.316	0.393	0.420	0.288	0.308	0.379
anni006	0.298	0.387	0.413	0.280	0.332	0.364
anni009	0.303	0.403	0.420	0.297	0.353	0.370
nad53	0.350	0.412	0.453	0.304	0.349	0.408
nad57	0.316	0.391	0.452	0.305	0.316	0.366

Table 5
Comparison of proposed method with other LBP based methods.

Video sequence	No. o shots	Original LBP (Ojala et al., 1996)			Midrange LBP (Rashmi, 2016)			Proposed Method		
		R	P	F	R	P	F	R	P	F
"anni005"	38	0.76	0.94	0.84	0.76	0.88	0.82	1.00	0.95	0.97
"anni006"	41	0.66	0.77	0.71	0.83	0.83	0.83	0.97	0.93	0.95
"anni009"	38	0.74	0.88	0.80	0.79	0.86	0.86	1.00	0.97	0.98
"nad53"	83	0.77	0.77	0.77	0.87	0.87	0.87	0.98	0.96	0.97
Average		0.73	0.84	0.78	0.81	0.86	0.84	0.98	0.95	0.97

Table 6
Comparison of proposed method with non-LBP SBD methods.

Video sequence	No. of shots	Chakraborty and Thounaojam (2020)			Rashmi and Nagendraswamy (2018)			Hannane et al., (2016)			Proposed Method		
		R	P	F	R	P	F	R	P	F	R	P	F
"anni005"	38	90.50	76.0	82.60	100	95	97.44	100	90.4	94.96	100	95	97.44
"anni009"	38	89.50	97.10	93.20	100	87.5	93.33	100	84.5	91.60	100	97.43	98.69
"nad53"	83	–	–	–	98.7	92.8	95.66	98.8	88.1	93.14	98.8	96.47	97.62
"nad57"	44	100	100	100	100	97.9	98.94	100	95.6	97.75	100	97.77	98.87
Average		93.33	91.03	91.93	99.67	93.3	96.34	99.7	89.65	94.36	99.7	96.66	98.15

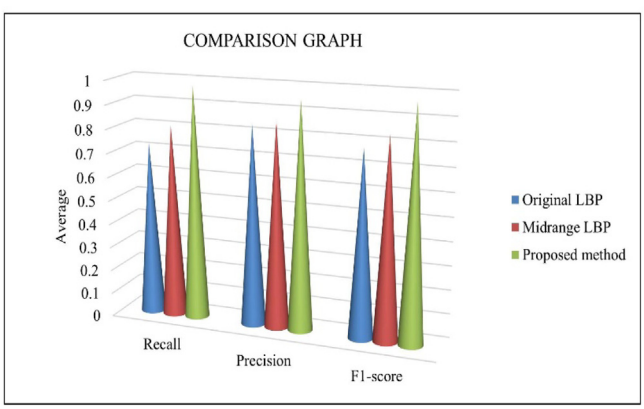


Fig. 7. Comparison graph of proposed method with LBP based methods.

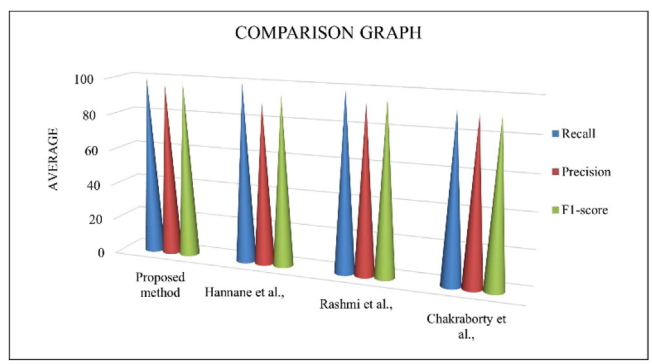


Fig. 8. Comparison graph of proposed method with non-LBP SBD methods.

sensitive to sudden illumination and camera panning/zooming operations (Fig. 9).

4.2. Performance evaluation for keyframe extraction

One of the important task in video analysis and summarization is the evaluation of extracted keyframes. Lack of ground truth makes evaluation of extracted keyframes a tedious task. The keyframe extracted from a shot must hold the maximum information conveyed through a shot. In the proposed method, fidelity measure which is based on Semi-Hausdorff distance is used to check the effectiveness of the extracted keyframes. The fidelity measure (Chang et al., 1999) is computed as the maximum of the minimum distances between the keyframe set and the shot frame set. Let V = $\{f1, f2, f3, \ldots fn\}$ be the n frames of input video and $R = \{kf1, kf2, kf3, \ldots kfm\}$ be the set of m keyframes extracted from shots of a video sequence. The distance between the sets are defined as follows:

$$d_j = \min(d(V_i, R_{km})) \tag{11}$$

Each frame of a video is characterized using the features as discussed earlier in feature extraction section. The Semi-Hausdorff distance between R and V is defined as:

$$d_{sh} = \max(d_j) \tag{12}$$

And, the fidelity measure is defined as:

$$fidelity = \frac{(1 - d_{sh})}{\max(\max(d_i))} \tag{13}$$

High fidelity value indicate that the extracted keyframe set provides a good global description of the visual content of the video sequence. The fidelity measure values of the proposed method,

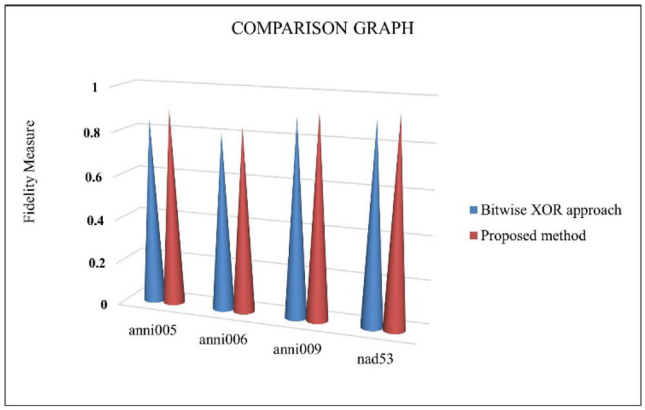


Fig. 9. Comparison graph in terms of fidelity.

Table 7
Comparative analysis in terms of fidelity.

Video Name	Abrupt cuts	Bitwise XOR approach(Rashmi and Nagendraswamy, 2016)	Proposed Method
anni005	38	0.85	0.91
anni006	41	0.81	0.84
anni009	38	0.90	0.93
NAD53	83	0.91	0.94
Average		0.86	0.90

implemented on TRECVID 2001 data set is recorded in Table 7 and comparative analysis of results has been performed with Bitwise-XOR approach (Rashmi and Nagendraswamy, 2016). The result shows that, the proposed method outperforms Bitwise-XOR method (Rashmi and Nagendraswamy, 2016) with 90% fidelity score. Extracted keyframes from a video shot of TRECVID 2001 dataset are good enough to represent and summarize the entire original video in a compact manner.

5. Conclusion

In this research work, a simple and efficient method to generate video summary is proposed, which aids for video indexing and retrieval. Proposed algorithm detects abrupt transitions by extracting binary edge information of the video frames for LBP histogram characterisation. The distance comparison of the adjacent frames is computed using Euclidean Distance and an adaptive threshold is employed to detect the abrupt cuts. The experimental analysis conducted for SBD on TRECVID 2001 dataset signifies that the BELBP histograms have better discriminative capability and provides good results with average F1-score of 98.15%. During keyframe extraction, Z-score is applied on the estimated magnitude of Sobel gradient images and co-efficient of variation is computed. A video summary is generated by combining the selected keyframes that possess the highest co-efficient of variation from every video shot. The chosen keyframe from the video shot is efficient enough to build a video summary with average fidelity measure of 90%. In future work, proposed method will be analyzed to detect the gradual transitions in the video and evaluation of other visual features to extract the keyframes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdesselam, A., 2013. Improving local binary patterns techniques by using edge information. Lect. Notes Softw. Eng. 360–363. https://doi.org/10.7763/lnse.2013.v1.77.

Abdulhussain, S.H., Ramli, A.R., Mahmmod, B.M., Saripan, M.I., Al-Haddad, S.A.R., Jassim, W.A., 2019. Shot boundary detection based on orthogonal polynomial. Multimed. Tools Appl. 78, 20361–20382. https://doi.org/10.1007/s11042-019-7364-3.

Abdulhussain, S.H., Ramli, A.R., Saripan, M.I., Mahmmod, B.M., Al-Haddad, S.A.R., Jassim, W.A., 2018. Methods and challenges in Shot boundary detection: a review. Entropy. https://doi.org/10.3390/e20040214.

Adjeroh, D., Lee, M.C., Banda, N., Kandaswamy, U., 2009. Adaptive edge-oriented shot boundary detection. Eurasip J. Image Video Process. 2009. https://doi.org/10.1155/2009/859371.

Angadi, S., Naik, V., 2014. Entropy based fuzzy C means clustering and key frame extraction for sports video summarization, in: Proceedings - 2014 5th International Conference on Signal and Image Processing, ICSIP 2014. IEEE Computer Society, pp. 271–279. https://doi.org/10.1109/ICSIP.2014.49

Badshah, N., Chen, K., Ali, H., Murtaza, G., 2012. Coefficient of variation based image selective segmentation model using active contours. East Asian J. Appl. Math. 2, 150–169. https://doi.org/10.4208/eajam.090312.080412a.

Bai, G., Zhu, Y., Ding, Z., 2008. A hierarchical face recognition method based on Local Binary Pattern. Proc. - 1st Int. Congr. Image Signal Process. CISP 2008 2, 610–614. https://doi.org/10.1109/CISP.2008.520

Barhoumi, W., Zagrouba, E., 2013. On-the-fly extraction of key frames for efficient video summarization. AASRI Procedia 4, 78–84. https://doi.org/10.1016/j.aasri.2013.10.013.

Besiris, D., Fotopoulou, F., Economou, G., Fotopoulos, S., 2008. Video Summarization by a Graph-Theoretic FCM based algorithm.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8, 679–698. https://doi.org/10.1109/TPAMI.1986.4767851.

Chakraborty, S., Thounaojam, D.M., 2020. SBD-Duo: a dual stage shot boundary detection technique robust to motion and illumination effect. Multimed. Tools Appl. https://doi.org/10.1007/s11042-020-09683-y.

Chang, H.S., Sull, S., Lee, S.U., Member, S., 1999. Efficient Video Indexing Scheme for Content-Based Retrieval 9, 1269–1279.

Dadashi, R., Kanan, H.R., 2013. AVCD-FRA: a novel solution to automatic video cut detection using fuzzy-rule-based approach. Comput. Vis. Image Underst. 117, 807–817. https://doi.org/10.1016/j.cviu.2013.03.002.

Duan, X., Lin, L., Chao, H., 2013. Discovering video shot categories by unsupervised stochastic graph partition. IEEE Trans. Multimed. 15, 167–180. https://doi.org/10.1109/TMM.2012.2225029.

El khattabi, Z., Tabii, Y., Benkaddour, A., 2017. Video shot boundary detection using the scale invariant feature transform and RGB color channels. Int. J. Electr. Comput. Eng. 7, 2565–2573. https://doi.org/10.11591/ijece.v7i5.pp2565-2573.

Ford, R.M., Robson, C., Temple, D., Gerlach, M., 2000. Metrics for shot boundary detection in digital video sequences. Multimed. Syst. 8, 37–46. https://doi.org/10.1007/s005300050003.

Furini, M., Geraci, F., Montangero, M., Pellegrini, M., 2010. STIMO: STIll and MOving video storyboard for the web scenario. Multimed. Tools Appl. 46, 47–69. https://doi.org/10.1007/s11042-009-0307-7.

Gharbi, H., Bahroun, S., Zagrouba, E., 2019. Key frame extraction for video summarization using local description and repeatability graph clustering. Signal, Image Video Process. 13, 507–515. https://doi.org/10.1007/s11760-018-1376-8.

Guru, D.S., Suhil, M., Lolika, P., 2013. A Novel Approach for Shot Boundary Detection in Videos.

Hannane, R., Elboushaki, A., Afdel, K., Naghabhushan, P., Javed, M., 2016. An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram. Int. J. Multimed. Inf. Retr. 5, 89–104. https://doi.org/10.1007/s13735-016-0095-6.

Huo, Y., Wang, Y., Hu, H., 2016. Effective algorithms for video shot and scene boundaries detection. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 – Proceedings. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICIS.2016.7550913.

Jadhav, P.S., Jadhav, D.S., 2015. Video Summarization using higher order color moments (VSUHCM), in: Procedia Computer Science. Elsevier B.V., pp. 275–281. https://doi.org/10.1016/j.procs.2015.03.140.

Jin, H., Liu, Q., Lu, H., Tong, X., 2004. Face detection using improved LBP under bayesian framework. Proc. - Third Int. Conf. Image Graph. 306–309. https://doi.org/10.1109/icig.2004.62.

Kanagaraj, K., Priya, G.G.L., 2018. Curvelet transform based feature extraction and selection for multimedia event classification. J. King Saud Univ. - Comput Inf. Sci. 0–8. https://doi.org/10.1016/j.jksuci.2018.11.006.

Khaleefah, S.H., Mostafa, S.A., Mustapha, A., Nasrudin, M.F., 2019. The ideal effect of Gabor filters and Uniform Local Binary Pattern combinations on deformed scanned paper images. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2019.07.012.

Liao, S., Law, M.W.K., Chung, A.C.S., 2009. Dominant local binary patterns for texture classification. IEEE Trans. Image Process. 18, 1107–1118. https://doi.org/10.1109/TIP.2009.2015682.

Mussel Cirne, M.V., Pedrini, H., 2018. VISCOM: a robust video summarization approach using color co-occurrence matrices. Multimed. Tools Appl. 77, 857–875. https://doi.org/10.1007/s11042-016-4300-7.

Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on feature distributions. Pattern Recognit. 29, 51–59. https://doi.org/10.1016/0031-3203(95)00067-4.

Pan, G., Zheng, Y., Zhang, R., Han, Z., Sun, D., Qu, X., 2019. A bottom-up summarization algorithm for videos in the wild. EURASIP J. Adv. Signal Process. 2019. https://doi.org/10.1186/s13634-019-0611-y.

Pietikainen, M., 2005. Texture analysis with local binary patterns. Handb. Pattern Recognit. Comput. Vis. https://doi.org/10.1142/9789812775320.

Qian, X., Hua, X.S., Chen, P., Ke, L., 2011. PLBP: an effective local binary patterns texture descriptor with pyramid representation. Pattern Recognit. 44, 2502–2515. https://doi.org/10.1016/j.patcog.2011.03.029.

Rashmi B. S., b N.H.S., DoS in Computer Science, DoS in Computer Science, University of Mysore, Mysore, I., University of Mysore, bKarnataka State Open University, Mysore, I., Mysore, I., Rashmibsrsh@compsci.uni-mysore.ac.in, 2016. Video Shot Boundary Detection using Midrange Local Binary Pattern. pp. 201–206.

Rashmi, B.S., Nagendraswamy, H.S., 2018. Effective video shot boundary detection and keyframe selection using soft computing techniques. Int. J. Comput. Vis. Image Process. 8, 27–48. https://doi.org/10.4018/ijcvip.2018040102.

Rashmi, B.S., Nagendraswamy, H.S., 2016. Shot-based keyframe extraction using bitwise-XOR dissimilarity approach. Int. Conf. Recent Trends Image Process. Pattern Recogn., 305–316 https://doi.org/10.1007/978-981-10-4859-3.

Roberts, L.G., 1963. Roberts 33959125-MIT.pdf.

Singh, A., Thounaojam, D.M., Chakraborty, S., 2020. A novel automatic shot boundary detection algorithm: robust to illumination and motion effect. Signal, Image Video Process. 14, 645–653. https://doi.org/10.1007/s11760-019-01593-3.

Sliti, O., Hamam, H., Amiri, H., 2018. CLBP for scale and orientation adaptive mean shift tracking. J. King Saud Univ. - Comput Inf. Sci. 30, 416–429. https://doi.org/10.1016/j.jksuci.2017.05.003.

Sobel, I., Feldman, G., 1968. An Isotropic 3x3 Image Gradient Operator. Stanford Artif. Intell. Proj., 271–272

Thakre, K.S., Rajurkar, A.M., Manthalkar, R.R., 2016. Video Partitioning and Secured Keyframe Extraction of MPEG Video, in: Physics Procedia. Elsevier B.V., pp. 790–798. https://doi.org/10.1016/j.procs.2016.02.058

Truong, B.T., Venkatesh, S., 2007. Video abstraction: a systematic review and classification. ACM Trans. Multimed. Comput. Commun. Appl. 3. https://doi.org/10.1145/1198302.1198305.

Yong, S.P., Deng, J.D., Purvis, M.K., 2013. Wildlife video key-frame extraction based on novelty detection in semantic context. Multimed. Tools Appl. 62, 359–376. https://doi.org/10.1007/s11042-011-0902-2.

Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., Zhang, B., 2007. A formal study of shot boundary detection. IEEE Trans. Circuits Syst. Video Technol. 17, 168–186. https://doi.org/10.1109/TCSVT.2006.888023.

Keyframe Extraction Using Sobel Fuzzified Weighted Approach

H. M. Nandini[1,2](✉), H. K. Chethan[1], and B. S. Rashmi[2]

[1] Department of Computer Science and Engineering, Maharaja Research Foundation, Maharaja Institute of Technology, Mysuru, India
[2] Department of Information Technology, Karnataka State Open University, Mysuru, India

Abstract. Nowadays, progress in technology and the application of internet has led to exponential growth of video data. This drastic increase stipulates efficient video analysis techniques. Keyframe extraction is one of the technique that provides a succinct representation of video and are useful in various applications like video indexing and retrieval. In this direction, an efficient approach for keyframe extraction is proposed. The process begins by converting gray scale images of shots into gradient magnitude images using Sobel operator and establishing fuzzification. Further, 3*3 mask of sliding window is utilized in both overlapping and non-overlapping fashion to obtain Binary Weighted Codes (BWC) on a fuzzified edge image. In the subsequent step, feature set is obtained by applying entropy measure on BWC values of every frame within a shot. Finally, frame having highest entropy value is chosen as a keyframe of the corresponding shot. To verify the effectiveness of the proposed approach experiments were conducted on Open Video Project dataset. The experimental results shows that the proposed method outperforms some of the state-of-the-art algorithms with average of 93% fidelity measure.

Keywords: Sobel descriptor · Gradient function · Keyframe extraction · Fuzzification · Entropy measure

1 Introduction

The advancement in various multimedia applications has driven digital video content as an emerging force. This rapid increase in video data demands efficient techniques for video summarization, indexing, browsing and retrieval. Therefore, there is an immense scope for researchers to solve the problem of video data analysis. The huge size of video content is an obstacle to many applications, and hence, there is a need of perspective frames which represents the entire video [1]. The approach of removing redundant frames from the video and obtaining its concise version is termed as video summarization [2]. Video summaries holds salient information of the video and avoids redundancy, but at the same time, preserves the original information of the video [3]. It facilitates browsing of large video database and complements content based video retrieval approach [4]. Still images (keyframes) and moving images (video skims) are the two most popular

A. Abraham et al. (Eds.): ISDA 2020, AISC 1351, pp. 236–246, 2021.
https://doi.org/10.1007/978-3-030-71187-0_22

ways to generate video summaries [5]. Advantage of video skim over static summary is the ability to retain dynamic mode of video information by including audio and motion elements that compliments the information to be conveyed by the summary [6]. In spite of these reasons, extracting of still images is still essential as it gives more flexibility without synchronization issues [7].

In literature, there are various techniques to extract keyframes with the justifiable quality but they are computationally expensive [8]. There are numerous ways to extract features from images: based on texture, shapes, edges etc. Edge is considered as an essential feature that represent the content of the image and several research works have been carried out using edge detectors [9, 10]. To address uncertainty in images, researchers have incorporated various fuzzy logic techniques in conjunction with edge descriptors. However, the results of the classical edge detection methods can be improved by transforming the gradient magnitude into suitable membership degrees to generate the fuzzy edge images [11]. This has motivated us to carry out the proposed research work. In this approach, videos are segmented into shots using algorithm proposed by Rashmi et al. [12]. Here, gray scale images of video shots are converted into gradient magnitude images using Sobel operator and fuzzification is established by incorporating triangular membership function. Further, Binary Weighted Codes (BWC) are generated by applying binary weights to 3*3 fuzzified image mask at every pixel position. Extensive experiments have been conducted considering each pixel in overlapping and non-overlapping mode. In order to extract frame feature, entropy measure is applied. Frames that possess maximum entropy value is selected as a keyframe of the corresponding shot and video summary is generated. The result obtained is evaluated using performance measure and compared with state-of-the-art approaches.

The rest of the paper is organized as follows: Sect. 2 gives the brief description of some related works. Section 3 presents the detailed description of the proposed methodology. Section 4 reports the experimental analysis and results and Sect. 5 concludes the paper.

2 Literature Review

Several methods have been proposed in the literature for keyframe extraction, as it plays a prerequisite role in video summarization, indexing and retrieval. An extensive review of the existing methods is found in [2, 13]. According to Angadi et al. [14] keyframe extraction techniques are categorized into four classes. The first category is sampling based method, where keyframes are extracted randomly without considering the content of the video. Though this method performs efficiently, sometimes fails to generate representative frames for short video clips. In object based method, keyframes extracted are efficient and semantic but it is suitable for certain applications only as it gives more importance to foreground. In segmentation based method, extracted keyframes efficiently represents the video content. However, segmentation is a complex process and deciding the number of segments is a hard task. In shot based method, keyframes are selected in an efficient way and it is considered as one of the important technique as it select keyframes from each shot.

A novel method for keyframe extraction was proposed by [15], where keyframes are extracted based on object detection. This algorithm employs combination of mutual

information entropy and SURF features. The authors in [16] used bitwise XOR variance method to extract the features of the video frames and keyframes are extracted based on the interframe dissimilarity matrix. Hannane et al. [17] have presented an efficient approach to select keyframes utilizing mean shift algorithm and global orientation feature. The approach proposed in [18] is based on fusion of Convolution Neural network based deep features and histogram. It generates the keyframe dynamically with less computational complexity. In [19], the authors have introduced an efficient approach to extract keyframes based on logical image description using interest points, repeatability network and modularity has been designed. The approach proposed in [20] used entropy based singular value metric to extract the keyframes from the segmented shots of the video. In [21], an efficient shot based keyframe extraction has been introduced. Here, for every frame multidimensional fuzzy histograms are constructed and keyframe is selected by minimizing cross-correlation criterion. Authors in [22] presented clustering based keyframe extraction based on energy rank obtained from dissimilarity and representativeness of video frames. Rashmi et al. [23] have utilized probabilistic entropy measure to choose the keyframe within fuzzified frames of video shot. The approach presented in [8] is based on fast clustering where keyframes are extracted using HSV color distribution of the frames. In [24], an online learnable module for keyframe selection has been developed. Locally-Consistent Deformable Convolution (LCDC) considering ResNet CNN are used to build the module.

3 Proposed Methodology

The proposed approach of keyframe extraction involves two phases viz. feature extraction and keyframe selection. The primary step is to convert the shot frames/images from color to gray scale before moving ahead to afore mentioned phases. Initially, gray scale images are transformed into gradient magnitude images using Sobel operator. Further, membership values are obtained by incorporating fuzzification process. Binary Weighted Codes (BWC) are computed by applying binary weights on every 3*3 image mask of fuzzified gradient magnitude frames. In the subsequent step, entropy measure is applied for BWC frame values. Finally, frame having highest entropy value within a shot is chosen as a keyframe and video summary is generated. Figure 1 depicts the general framework of the proposed methodology.

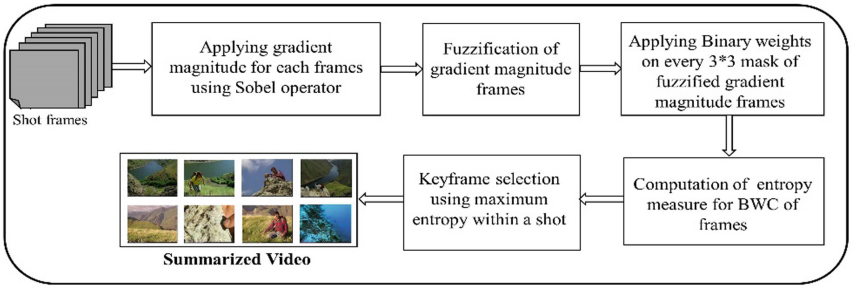


Fig. 1. Framework of the proposed methodology

3.1 Feature Extraction and Representation

Features that represents the visual information plays a vital role in identifying the keyframe within a video shot. In this approach, feature extraction is performed at three steps namely: (i) Conversion of gray scale images to Sobel gradient magnitude images (ii) Fuzzification of Sobel gradient magnitude images (iii) Computation of BWC and applying entropy measure. The following subsection describes each step in detail.

Conversion of Gray Scale Images to Sobel Gradient Magnitude Images
Edge detection aids to remove the irrelevant information and retains essential structural properties of the image [25]. In literature, there are various edge descriptors to obtain edge information. However, most of the approaches are based on gradient [11], based on the observation considering Canny [26], Sobel [27], Roberts [28] and some other recent methods [29, 30]. The gradient magnitude plays an important role in further processing steps. It can be utilized as the only information for subsequent steps [31] or in combination with the orientation of vector [26]. In the proposed method, Sobel edge descriptor is chosen to transform gray scale image to gradient magnitude image as it outperforms other edge detectors in terms of computational efficiency and accuracy [9]. The Sobel descriptor consists of a pair of 3*3 kernels which are convolved with original image to compute approximations of the derivatives, one estimating the gradient in x-direction (columns) and the other in y-direction (rows) as shown in equation.

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad G_y = \begin{bmatrix} +1 & 2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{1}$$

Kernels produces separate measurements for the gradient component in each direction, namely G_{x} and G_{y}. The gradient magnitude is computed using following equation.

$$GM = \sqrt{G_x^2 + G_y^2} \tag{2}$$

The illustration of Sobel gradient magnitude image for a sample image taken from the dataset is depicted in Fig. 2.

Fuzzification of Sobel Gradient Magnitude Images
Fuzzification of an image indicates the transformation of image data from gray level values [0, 255] to real values [0, 1] using Membership Function (MF) [32]. Utilization of fuzzy reasoning technique with sobel edge detector enhances the edge feature of the image [33]. In literature, triangular membership function is the most typical choice to represent information by means of fuzzy sets [11]. In this method, triangular membership function is utilized to estimate the fuzzified/membership value of each pixel of Sobel gradient magnitude image. For an image M, a fuzzy image MFS in a finite set X is formulated as:

$$M_{FS} = \left\{ b_{ij}, \mu_M(b_{ij}) | b_{ij} \in X \right\} \tag{3}$$

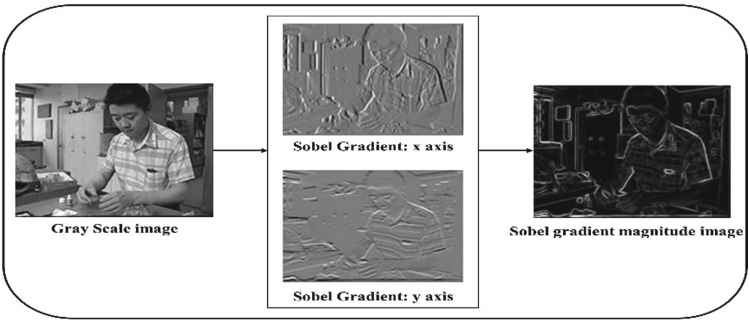


Fig. 2. Illustration of sobel gradient magnitude for frame #3096 of UGS01_004 video.

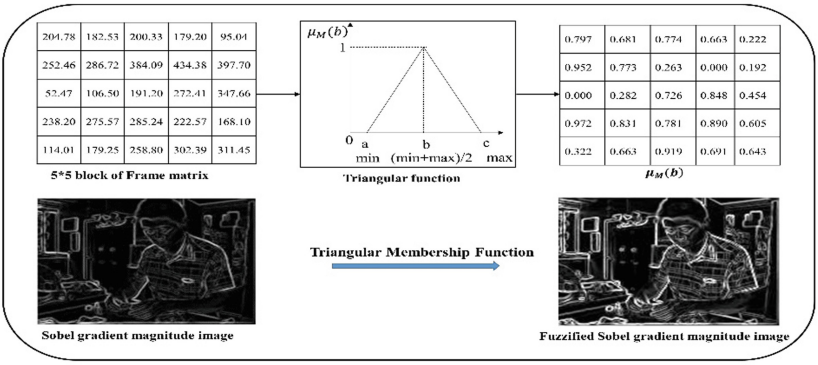


Fig. 3. Illustration of fuzzified sobel gradient magnitude image for frame #3096 of UGS01_004 video.

Where b_{ij} represents the pixel value of the Sobel gradient magnitude image and the function $\mu_M\ (b) \in [0, 1]$ is the fuzzified value. Figure 3 shows the fuzzification of the image using fuzzy triangular MF.

Computation of BWC and Applying Entropy Information Measure

This sub section describes, feature extraction process using binary weights to 3*3 image mask and computing entropy measures for keyframe extraction. Here, fuzzified values are further processed using 3*3 sliding window mask at each pixel position. The evaluation process is initiated from left to right and top to bottom at respective pixel positions both in overlapping and non-overlapping fashion.

BWC is generated for each sliding window mask by traversing in clockwise direction as depicted in the following equation.

$$BWC = \sum_{i=0}^{8} (fp)2^i \tag{4}$$

Where *fp* represents the fuzzified pixel values of an image. The code values generated for each 3*3 mask in an overlapping fashion is illustrated in Fig. 4.

After obtaining BWC from the above mentioned process for all the frames of the shot, entropy measure is computed as depicted in the following equation

$$En = -\sum_{k=0}^{n-1} (BWC_k * \log_2(BWC_k)) \tag{5}$$

Where k represents the number of BWC levels and BWC_k is the probability distribution of BWC values obtained for Fuzzified Sobel gradient magnitude frame.

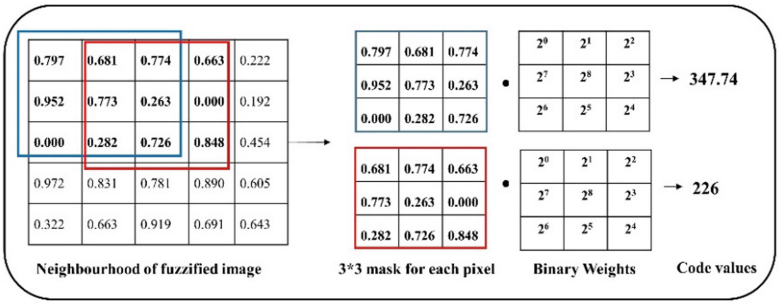


Fig. 4. Illustration of BWC values for 3*3 mask of an image

3.2 Keyframe Extraction

A keyframe is a representative frame which efficiently reflects the prominent content of the video shot. This keyframe aids to reduce the data size of video index and provides a systematized structure for the video stream. Here, entropy values computed in the Sect. 3.1 is considered to extract the keyframes from every shot. The entropy value is always non-negative and a variable with higher entropy value contains maximum information [34]. Hence, a frame possessing the highest entropy value amongst the computed entropies within a shot is chosen as a keyframe.

Let a video shot Sh consists of m frame sequences, then the computed entropy feature set is $Sh = \{En(f_1), En(f_2), En(f_3), En(f_4), \ldots\ldots En(f_m)\}$ where $En(f_i)$ corresponds to entropy feature value of ith frame within a shot. In order to select the keyframe for a shot Sh, the frame with the highest entropy value is identified as shown in the equation

$$keyf = \max(En(f_1), En(f_2), En(f_3), En(f_4), \ldots\ldots\ldots En(f_m)) \tag{6}$$

In this manner, the keyframes are extracted from every shot of the video and are concatenated to build a video summary.

4 Results and Discussion

Dataset: In order to analyze the performance of the proposed method, video sequences from Open Video Project dataset has been considered and described in Table 1. Extensive experiments were conducted to analyze the performance of the proposed technique using Matlab 2014a on Intel Core i5 processor, running at 2.70 GHz with 8 GB RAM.

Table 1. Description of open video project dataset

File name	Video name	Duration	Dimension	Total frames	No. of shots
UGS01_004	Exotic terrane, Segment 4	2:40	352×240	4,797	23
UGS02_009	America's new frontier, Segment 9	3:49	352×240	6,879	20
UGS03_004	The future of energy gases, Segment 4	4:27	352×240	8,007	27
UGS07_003	Ocean floor legacy, Segment 3	2:38	352×240	4,749	17
UGS07_005	Ocean floor legacy, Segment 5	2:35	352×240	4,665	25

4.1 Performance Evaluation of Keyframe Extraction

Evaluation of the extracted keyframe is one of the essential task in video analysis and summarization. Lack of ground truth makes evaluation of the extracted keyframes a tedious task. In the proposed method, keyframe extraction is performed using shot-based approach to construct a video summary. Thus, one frame per video shot is extracted as a keyframe. If the keyframe extracted exhibits the prominent visual contents of the entire shot, then the video summary is efficient. Fidelity measure based on Semi-Hausdorff distance is used to evaluate the effectiveness of the extracted keyframes. The fidelity measure [35] is computed as the maximum of the minimum distances between the keyframe set and the shot frame set. Let $G = \{f_1, f_2, f_3, \ldots f_n\}$ be the n frames of input video and $C = \{kf_1, kf_2, kf_3, \ldots kf_m\}$ be the set of m keyframes extracted from shots of a video sequence. The distance between the sets are defined as follows:

$$d_j = \min(d(G_i, C_{km})) \tag{7}$$

Each frame of a video is characterized using the features as discussed earlier in feature extraction section. The Semi-Hausdorff distance between G and C is defined as:

$$d_{sh} = \max(d_j) \tag{8}$$

And, the fidelity measure is formulated as:

$$Fidelity = \frac{(1 - d_{sh})}{\max(\max(d_i))} \tag{9}$$

High fidelity value indicate that, the extracted keyframe set is accurate and provides a good global description of the visual content of the video sequence.

Systematic experimentations have been carried out considering sliding process of respective frame pixels. The results obtained by the proposed approach implemented on Open Video Project datasets are recorded in Table 2 considering 3*3 mask for the image in overlapping and non-overlapping fashion.

Table 2. Results of non-overlapping and overlapping methods in terms of fidelity measure

Video sequence	Non-overlapping method	Overlapping method
UGS01_004	0.87	0.92
UGS02_009	0.91	0.94
UGS03_004	0.89	0.93
UGS07_003	0.86	0.90
UGS07_005	0.90	0.94
Average	0.89	0.93

Significance of overlapping and non-overlapping method of sliding window has been analyzed. Obtained results exhibits that, the performance of overlapping method considering 3*3 mask is better than non-overlapping method in terms of fidelity measure. However, the time complexity of the overlapping method slightly exceeds non-overlapping method.

The best case results discussed in terms of fidelity measure is compared with other state-of-the-art algorithms [23, 36] and are recorded in Table 3 and the performance is represented graphically in Fig. 5.

Table 3. Comparative analysis in terms of fidelity measure.

Video sequence	Besiris et al. [36]	Rashmi et al. [23]	Proposed method
UGS01_004	0.72	0.89	0.92
UGS02_009	0.73	0.93	0.94
UGS03_004	0.73	0.91	0.93
UGS07_003	0.75	0.89	0.90
UGS07_005	0.74	0.92	0.94
Average	0.73	0.91	0.93

The results portrays that, the proposed method of keyframe extraction outperforms some of the state of the art algorithms in terms of fidelity measure. Therefore the proposed approach to select keyframe from the video shot is efficient enough to build a video summary of the original video in a concise manner.

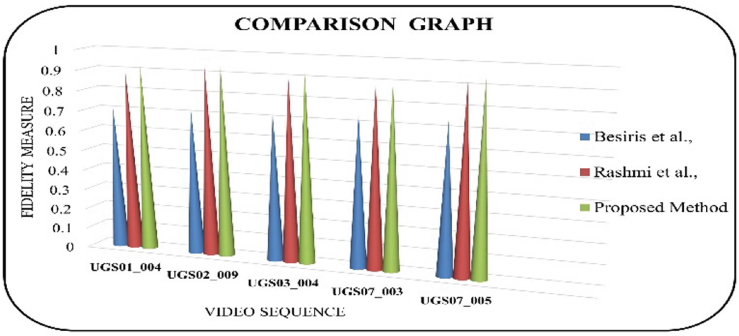


Fig. 5. Comparison graph of proposed method with other methods

5 Conclusion

In this research work, we have proposed a method to extract keyframes from shots of the video. Our approach extracts keyframes by exploiting Sobel gradient magnitude images and incorporating the concepts of fuzzy sets. Binary Weighted Codes are computed by utilizing sliding window of 3*3 mask on fuzzified Sobel gradient magnitude images. Entropy measure has been applied on BWC of every frame and frames that possess the highest entropy value from every video shot is selected as a keyframe to build a video summary. Experimental analysis and evaluation of the proposed method has been carried out on benchmark video datasets taken from Open Video Project. The overlapping method yields more accuracy than non-overlapping method. The proposed approach outperforms some of the state-of-the-art methods with 93% of average fidelity measure. In future work, performance of the proposed method will be investigated considering different membership functions to fuzzify the edge images.

References

1. Asha Paul, M.K., Kavitha, J., Jansi Rani, P.A.: Key-frame extraction techniques: a review. Recent Pat. Comput. Sci. **11**(1), 3–16 (2018)
2. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Representation **19**(2), 121–143 (2008)
3. De Avila, S.E.F., Lopes, A.P.B., Da Luz, A., De Albuquerque Araújo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognit. Lett. **32**(1), 56–68 (2011)
4. Hanjalic, A., Zhang, H.: An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Trans. Circuits Syst. Video Technol. **9**(8), 1280–1289 (1999)
5. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **3**(1), 3 (2007)
6. Jadhav, P.S., Jadhav, D.S.: Video summarization using higher order color moments (VSUHCM). In: Procedia Computer Science. Elsevier B.V., pp. 275–281 (2015)
7. Ejaz, N., Tariq, T.B., Baik, S.W.: Adaptive key frame extraction for video summarization using an aggregation mechanism. J. Vis. Commun. Image Representation **23**(7), 1031–1040 (2012)

8. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: STIMO: STIll and MOving video storyboard for the web scenario. Multimedia Tools Appl. **46**(1), 47–69 (2010)
9. Abdesselam, A.: Improving local binary patterns techniques by using edge information. Lect. Notes Softw. Eng. **1**(4), 360–363 (2013)
10. Nandini, H.M., Chethan, H.K., Rashmi, B.S.: Shot based keyframe extraction using edge-LBP approach. Journal of King Saud University - Computer and Information Sciences (2020)
11. Lopez-Molina, C., De Baets, B., Bustince, H.: Generating fuzzy edge images from gradient magnitudes. Comput. Vis. Image Underst. **115**(11), 1571–1580 (2011)
12. Rashmi, B.S., Nagendraswamy, H.S.: Video shot boundary detection using block based cumulative approach. Multimedia Tools and Applications, pp. 1–24 (2020)
13. Mashtalir, S., Mikhnova, O.: Key frame extraction from video: framework and advances. Int. J. Comput. Vis. Image Process. (IJCVIP) **4**(2), 68–79 (2014)
14. Angadi, S., Naik, V.: Entropy based fuzzy C means clustering and key frame extraction for sports video summarization. In: Proceedings - 2014, ICSIP 2014, pp. 271–279. IEEE Computer Society (2014)
15. Chen, M., Han, X., Zhang, H., Lin, G., Kamruzzaman, M.M.: Quality-guided key frames selection from video stream based on object detection. J. Vis. Commun. Image Representation **65**, 102678 (2019)
16. Rashmi, B.S., Nagendraswamy, H.S.: Shot-based keyframe extraction using bitwise-XOR dissimilarity approach. In: In International Conference on Recent Trends in Image Processing and Pattern Recognition, pp. 305–316. Springer, Singapore (2016)
17. Hannane, R., Elboushaki, A., Afdel, K.: MSKVS: adaptive mean shift-based keyframe extraction for video summarization and a new objective verification approach. J. Vis. Commun. Image Representation **55**, 179–200 (2018)
18. Ujwalla, G., Hajari, K., Yogesh, G.: Deep learning approach to key frame detection in human action action videos. In: Recent Trends in Computational Intelligence, vol. 13. IntechOpen (2020)
19. Gharbi, H., Bahroun, S., Zagrouba, E.: Key frame extraction for video summarization using local description and repeatability graph clustering. Sign. Image Video Process. **13**(3), 507–515 (2019)
20. Hannane, R., Elboushaki, A., Afdel, K., Naghabhushan, P., Javed, M.: An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram. Int. J. Multimedia Inform. Retrieval **5**(2), 89–104 (2016)
21. Doulamis, A.D., Doulamis, N.D., Kollias, S.D.: Efficient video summarization based on a fuzzy video content representation. In: 2000 IEEE (ISCAS), vol. 4, pp. 301–304 (2000)
22. Pan, G., Zheng, Y., Zhang, R., Han, Z., Sun, D., Qu, X.: A bottom-up summarization algorithm for videos in the wild. EURASIP J. Adv. Sign. Process. **2019**, 15 (2019)
23. Rashmi, B.S., Nagendraswamy, H.S.: Effective video shot boundary detection and keyframe selection using soft computing techniques. (IJCVIP) **8**(2), 27–48 (2018)
24. Elahi, G.M., Yang, Y.H.: Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition, vol. 12434 (2020). arXiv preprint arXiv:2009
25. Dhagdi, S.T., Deshmukh, P.R.: Keyframe based video summarization using automatic threshold & edge matching rate. Int. J. Sci. Res. Publ. **2**(7), 1–12 (2012)
26. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8, **6**, 679–698 (1986)
27. Sobel, I., Feldman, G.: An isotropic 3×3 image gradient operator. Stanford Artificial Intelligence Project, pp. 271–272 (1968)
28. Roberts, L.G.: Machine perception of three-dimensional solids. PhD diss., Massachusetts Institute of Technology (1963)
29. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 117–156 (1998)

30. Sun, G., Liu, Q.H., Liu, Q., Ji, C., Li, X.: A novel approach for edge detection based on the theory of universal gravity. Pattern Recogn. **40**(10), 2766–2775 (2007)
31. Rosin, P.L.: Unimodal thresholding. Pattern Recogn. **34**(11), 2083–2096 (2001)
32. Zadeh, L.A.: Fuzzy sets. Inform. Control **8**(3), 338–353 (1965)
33. Kuo, Y.H., Lee, C.S., Liu, C.C.: New fuzzy edge detection method for image enhancement. In: Proceedings of 6th International Fuzzy Systems Conference, vol. 2, pp. 1069–1074. IEEE (1997)
34. Wang, C., Shen, H.W.: Information theory in scientific visualization. Entropy **13**(1), 254–273 (2011)
35. Chang, H.S., Sull, S., Lee, S.U., Member, S.: Efficient video indexing scheme for content-based retrieval. IEEE Trans. Circ. Syst. Video Technol. **9**, 1269–1279 (1999)
36. Besiris, D., Fotopoulou, F., Laskaris, N., Economou, G.: Key frame extraction in video sequences: a vantage points approach. In: 2007 IEEE 9th International Workshopon Multimed Signal Process MMSP 2007, pp. 434–437. IEEE (2007)

International Journal of Computer Vision and Image Processing
Volume 11 • Issue 3 • July-September 2021

An Efficient Method for Video Shot Transition Detection Using Probability Binary Weight Approach

Nandini H. M., Maharaja Research Foundation, University of Mysore, India
Chethan H. K., Maharaja Research Foundation, University of Mysore, India
Rashmi B. S., Karnataka State Open University, India

ABSTRACT

Shot boundary detection in videos is one of the most fundamental tasks towards content-based video retrieval and analysis. In this aspect, an efficient approach to detect abrupt and gradual transition in videos is presented. The proposed method detects the shot boundaries in videos by extracting block-based mean probability binary weight (MPBW) histogram from the normalized Kirsch magnitude frames as an amalgamation of local and global features. Abrupt transitions in videos are detected by utilizing the distance measure between consecutive MPBW histograms and employing an adaptive threshold. In the subsequent step, co-efficient of mean deviation and variance statistical measure is applied on MPBW histograms to detect gradual transitions in the video. Experiments were conducted on TRECVID 2001 and 2007 datasets to analyse and validate the proposed method. Experimental result shows significant improvement of the proposed SBD approach over some of the state-of-the-art algorithms in terms of recall, precision, and F1-score.

KEYWORDS

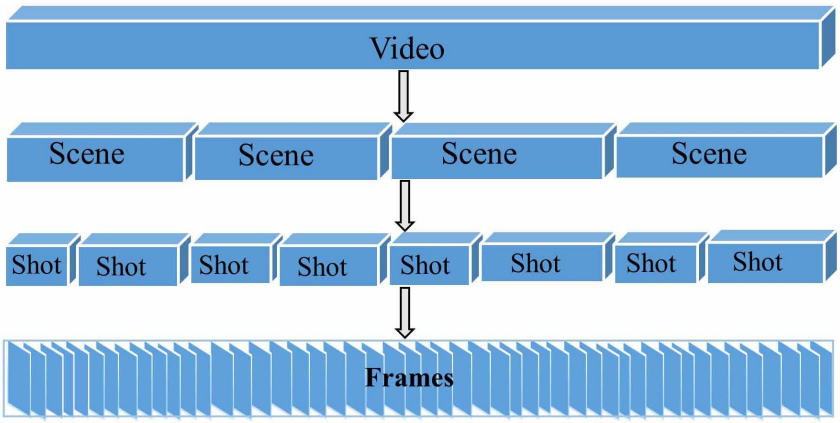
Co-Efficient of Mean Deviation, Gradual Transition, Histogram, Kirsch Operator, Normalization, Shot Boundary Detection, Variance, Z-Score

1. INTRODUCTION

The enormous development of multimedia technologies and availability of computing resources have led to the explosion of video data on the internet. However, the surge of video data has not been induced by an increase in its accessibility. Thus, there is a need for techniques that can efficiently access, browse, index and retrieve the video data. Shot Boundary Detection (SBD) is a preliminary step for video abstraction, video segmentation and video retrieval approaches (Hanjalic, 2002). A video shot represents a sequence of interrelated frames captured in a single take with one camera (Pal et al., 2015). Detection of shot boundaries in a video is mainly based on identifying the editing effects that are used to combine shots into video sequence. The hierarchal structure of video is represented in Figure 1.

DOI: 10.4018/IJCVIP.2021070101

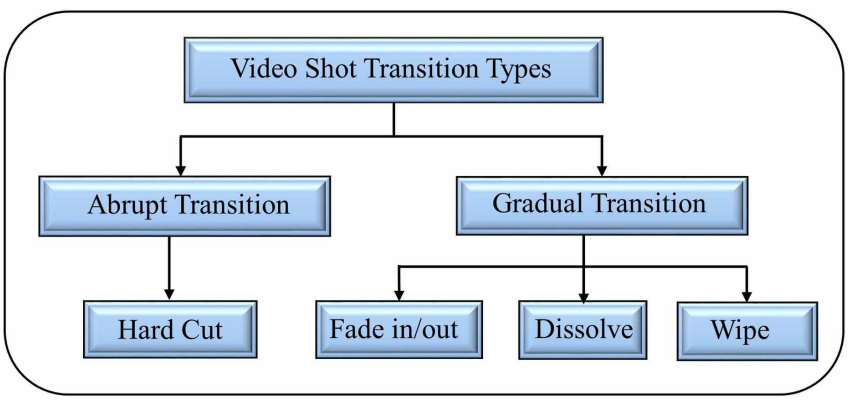
Figure 1. Hierarchal structure of video



Generally, transition of shots in a video are categorized as: Abrupt and Gradual transition (Sengupta et al., 2015) and it is depicted in Figure 2. Abrupt transition occurs when there is a rapid change between consecutive frames/images whereas, gradual transition occurs when change of the boundary is over multiple frames. Fade-in, Fade-out, dissolve and wipes are the frequently used editing effects of gradual transition.

Extraction of essential features from video frames that efficiently represents the visual information plays an important role in SBD (Jadon et al., 2001). The frame features can be extracted either globally or locally (Thounaojam et al., 2016) and it provides different information of frame at computational level. The global features describes the visual content of whole image and it is represented by a single vector. Contrastingly from global features, local features describes the visual content of the frame in patches or considering pixels of small group and it is represented by a set of vectors. However, global features have certain limitations such as scaling, sensitivity to noise, illumination variation and it often fails to identify the essential features of the image (Kabbai et al., 2018). Thus, global features are not suitable for few applications. Their flaws are fixed by the use of local features which encodes the local information to get the finest details of the image such as interest points (Kabbai et al., 2018). Hence, several approaches available in the literature have combined both local and global features in various domains such as image retrieval (Chaudhary & Upadhyay, 2014), SBD (Rashmi

Figure 2. Types of video shot transition



& Nagendraswamy, 2020), object detection (Muralidharan & Chandrasekar, 2012) etc., and obtained good results. This has motivated us to carry out the proposed SBD approach.

The main focus of the proposed approach is to detect abrupt and gradual transitions in video using combination of local and global features. The challenging task of SBD is to differentiate between the scene breaks and other changes like object motion, illumination and camera motion (Thounaojam et al., 2014). Edges are extensively invariant under local illumination changes and are considerably less affected by possible motion in the video (Adjeroh et al., 2009). Thus, primarily gray scale frames of the video are transformed into magnitude frames using Kirsch edge operator. The obtained edge magnitude image values are normalised by utilizing Z-score measure. The significance of probability of image pixels in its local/spatial neighbourhood aids in removal of noise (Zhang et al., 2013). Influenced by this, authors have explored probabilities of pixels in its adjacent neighbourhood by integrating it with the binary weights. Thus, block based probability binary weight method on each 3*3 block pixels of normalised magnitude frame has been employed. This local feature has the better discrimination capability and is robust to noise and illumination variation. Further, mean of the probability binary weights is computed and is used to produce Mean Probability Binary Weight (MPBW) histogram for every frame of the video. Each MPBW histogram describes the visual information of the video frame globally and collectively represents the whole video. The distance between each adjacent MPBW histogram of video frames is computed utilizing Euclidean Distance and an adaptive threshold is applied to detect the abrupt transitions in video. To detect the gradual transitions in video, Co-efficient of Mean Deviation (CMD) and variance statistical measure is applied on each MPBW histogram and adaptive threshold strategy is devised. The effectiveness of the proposed approach is verified in terms of recall, precision and F1-score by conducting experiments on TRECVID 2001 and 2007 benchmark datasets.

The rest of the paper is organized as follows: Section 2 provides the brief description of some related works reported in SBD. Section 3 presents the detailed description of the proposed methodology for feature extraction and SBD. Section 4 reports the experimental analysis and results and Section 5 concludes the paper.

2. LITERATURE REVIEW

In this section, a brief outline of literature on SBD approaches by various researchers is presented. Video SBD is a preliminary step for many video analysis application (Bhaumik, Bhattacharyya, et al., 2016). A comprehensive review of existing SBD approaches has been reported in (Singh & Aggarwal, 2015)(Pal et al., 2015). There exists a numerous variety of SBD approaches in literature which mainly differs by the kind of features used and the time required to perform computation. However, most of the approaches have explored pixel, histogram, edge, statistical and block based features to detect the transitions in video.

A new Pyramidal Opponent Color Shape (POCS) model to detect the abrupt transition and gradual transition has been presented (Sasithradevi & Mohamed Mansoor Roomi, 2020). This model uses shape and color features of opponent color space to segment videos using Bagged Trees Classifier (BTC). Motion based SBD utilising curvlet features has been presented (Kanagaraj & Priya, 2018) to segment videos into shots. These curvlet features exhibits object motion in various magnitude and orientations. An efficient approach to detect abrupt transitions in video has been presented (Nandini et al., 2020) using ELBP histograms. Here, binarized edge information of frames has been utilized for LBP texture characterization. Rashmi and Nagendraswamy, (2020) have used combination of both global and local features to identify abrupt and gradual transitions in video. In this technique, fuzzified Sobel gradient images are used to construct block based MCSH features to detect shot boundaries. In (Chakraborty et al., 2020), CIDE 2000 colour difference and mean luminance pattern has been utilized to detect abrupt and gradual transitions. This method efficiently tackles object motion and lighting effects of a video. Guru and Suhil, (2013) have introduced a non-parametric method for SBD in videos

with split and merge framework by utilizing color histograms. A dual stage SBD method has been introduced (Chakraborty & Thounaojam, 2020) for abrupt and gradual transition detection in video. In this approach, gradient similarity and luminance distortion of frames are used as a feature and two adaptive threshold to detect video transitions. Shekar and Uma, (2015) have explored maximum gradient value amongst eight orientations of Kirsch edge descriptor to detect shot transition in video. The first and second order moments are computed for gradient values of frames to detect abrupt transition. In (Thounaojam et al., 2017), Gist and local descriptor are utilized to detect shot transitions in video. Gist is a scale invariant feature which is used to get perceptual and semantic information of scene. This technique decreases computational complexity by processing at shot transition regions only. A new approach for detection of shot transition in videos has been introduced (Priya & Domnic, 2012). This method works by extracting edge strength utilizing orthogonal vectors from blocks of frame. A dual-detection method has been proposed (Jiang et al., 2013) for video transition detection based on human visual features using uneven blocked color histogram and pixel value difference. In (Janwe & Bhoyar, 2013), Just Noticeable Difference (JND) color histogram and adaptive threshold are utilized for SBD in video. A fuzzy rule based video scene cut detection approach has been presented (Dadashi & Kanan, 2013) which incorporates spatial and temporal features of video frames. An approach to detect abrupt shot cuts in video has been introduced (Rashmi & Nagendraswamy, 2016) by exploiting edge information. This method constructs histogram for each frame using 2*2 mask of sliding window in overlapping and non-overlapping fashion to assign binary weights to the extracted edge information of frames. Hannane et al., (2016), have presented a novel approach for SBD by extracting SIFT point distribution histograms from every frames of video. An efficient algorithm to detect shot boundaries in video has been presented (Rashmi B S & Nagendraswamy H S, 2018) using soft computing techniques by incorporating fuzzy and intuitionistic fuzzy sets.

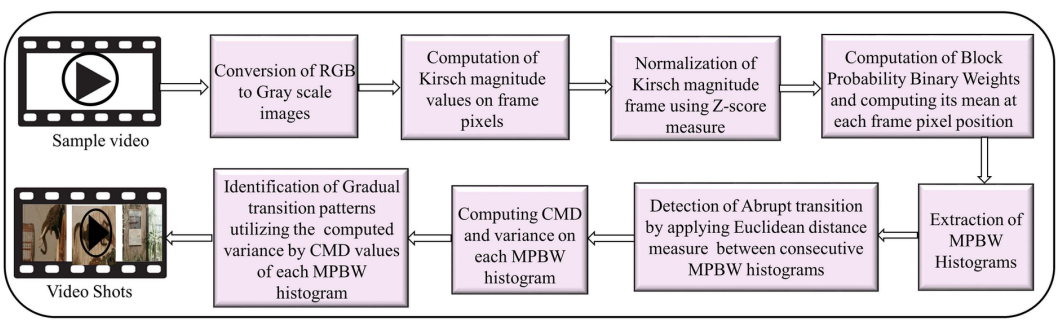
3. PROPOSED METHODOLOGY

In this section, techniques to detect abrupt and gradual transitions in video by using combination of local and global features has been discussed. The proposed approach consists of two phases namely, (i) Feature Extraction and Representation and (ii) Shot Boundary Detection. The framework of the proposed SBD method is illustrated in Figure 3. The initial step is to transform the extracted video frames from color to gray scale before moving forward to afore mentioned phases.

3.1 Feature Extraction and Representation

The accomplishment of any video SBD technique is highly relied on efficient visual content representation of the video frames. Thus, it is necessary to extract proficient features from the frames for detection of abrupt and gradual transitions in the video. Gray scale images are usually preferred

Figure 3. Framework of the proposed SBD approach



when compared to color images to extract descriptors as it simplifies the algorithm and decreases computational requirements. Algorithms applied to gray scale images are less sensitive to illumination effect and shows enhanced performance even if the illumination is variable (Kanan & Cottrell, 2012). Thus, in the proposed approach gray scale images are considered rather than color images to extract descriptors which doesn't affect the overall performance of the algorithm. The following subsections describes the process of feature extraction and representation of video frames.

3.1.1 Kirsch Magnitude Frames

The edge feature preserves the essential structural properties of an image and reduces the amount of data to be processed (Shekar & Uma, 2015). In literature, there exists immense variety of edge descriptors to measure the intensity changes such as Canny (Canny, 1986), Kirsch (Kirsch, 1971), Roberts (Roberts, 1963), Sobel (Sobel & Feldman, 1968) etc. However, Kirsch edge operator is one of the important edge descriptor that attains the good trade-off between retaining the edge information and suppressing the noise factor (Venmathi et al., 2016).

In the proposed method, Kirsch operator (Kirsch, 1971) is used to detect the maximum edge strength in the predetermined eight directions at every pixel position of an image. The edge direction is defined by single kernel mask which takes 45 degree rotation and increments through eight compass directions. The directions corresponds to west (W), southwest (SW), south (S), southeast (SE), east (E), northeast (NE), north (N), northwest (NW) and are represented as follows:

$$e_0 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} e_1 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} e_2 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} e_3 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \tag{1}$$

$$e_4 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} e_5 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix} e_6 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} e_7 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix}$$

Every frame of a video is convolved with kernel masks in eight compass directions $D_e(i, j)$ to get eight gradient images G_e as formulated in the equation:

$$G_e(i,j) = I(i,j) * D_e(i,j) \tag{2}$$

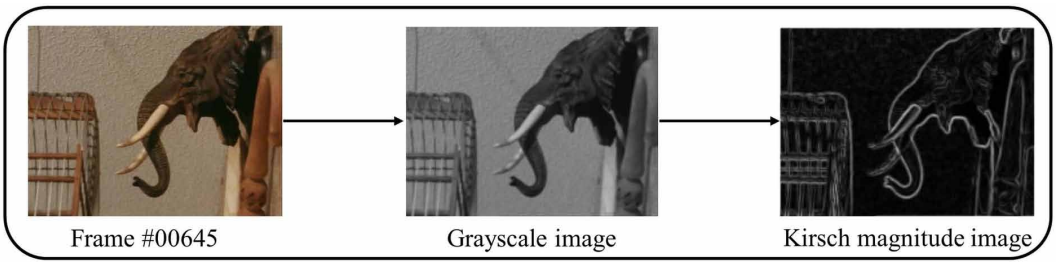
where $e \in [0\ldots7]$ represents eight direction. $I(i,j)$ represents the gray values of a pixel and its 3*3 eight neighbourhoods. The kirsch operator then extracts the maximum magnitude amongst eight gradient images to obtain the best direction information as formulated in the equation:

$$G(i,j) = \max_{e=0}^{7} G_e(i,j) \tag{3}$$

The illustration of Kirsch magnitude image for a sample image taken from TRCVID 2007 dataset is illustrated in Figure 4.

It is observed that increased kernel size may help to suppress the noise but it will deteriorate edge localization issues and forms interference between adjacent edges (A. et al., 2009). Thus, ideal 3*3 Kernel size is retained in all the experiments.

Figure 4. Illustration of Kirsch magnitude image for frame #00645 of BG_3097 video



3.1.2 Normalization of Kirsch Magnitude Frames

In quantitative image analysis, intensity normalization is essential, particularly while extracting the features of the image based on intensity (Sintorn et al., 2010). The Z-score transformation is one of the most commonly used statistical measure to normalise the data in neuroimaging (Ishii et al., 2000). In the proposed method, Z-score transformation is used to normalise the obtained edge magnitude image values. Z-score measures the position of every pixel of a frame in terms of standard deviation from the mean as formulated in the following equation and Figure 5 shows the illustration of computing the Z-score for Kirsch magnitude values:

$$ZG = (G_{ij} - \mu) / \sigma \tag{4}$$

where:

$$\mu = \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} G_{ij}\right)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(G_{ij} - \mu)^2}{n}}$$

3.1.3 Histogram Construction Using Block Based Probability With Binary Weights

The SBD method proposed in this work is based on the extraction of MPBW histogram features from the video frames. Here, normalised values obtained in sub-section 3.1.2 are further processed

Figure 5. Illustration of normalized Kirsch magnitude values for 5*5 pixel values

156	116	52	222	186
206	83	132	190	160
256	238	128	45	92
126	114	163	131	132
199	174	102	79	85

Kirsch magnitude pixel values

Z-score Transformation →

0.235	0.471	1.603	1.403	0.766
1.120	1.055	0.188	0.837	0.306
2.004	1.686	0.259	1.727	0.896
0.295	0.507	0.359	0.206	0.188
0.996	0.554	0.719	1.126	1.020

Normalized Kirsch magnitude pixel values

using 3*3 block at each pixel position. Probability Binary Weight Code (PBWC) values are obtained for each pixel of a block by applying the probability function with binary weights. The process of evaluation starts from left to right and top to bottom at respective pixel position in an overlapping fashion. PBWC values are obtained for each sliding window mask by traversing in clockwise direction as shown in the following equation:

$$P = ZG_{ij} \,/\, \sum_{i=1}^{3}\sum_{j=1}^{3} ZG_{ij} \tag{5}$$

$$PBW = P_{ij} * 2^{k} \tag{6}$$

where, {K=0 to 8}, *PBW* represents the PBWC values and *P* is the probability pixel values of a 3*3 block. Further, mean of PBWC values for each block is formulated as follows:

$$\mu = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} PBW_{ij}}{3*3} \tag{7}$$

Here, μ represents the mean value of PBWC values for each 3*3 block in an overlapping fashion. Thus, histogram is constructed utilizing the obtained mean value for every 3*3 block at each pixel position which helps in the detection of shot boundary. Figure 6 shows the illustration of computing the Mean Probability Binary Weight (MPBW) values for 3*3 block of a sample frame and Figure 7 represents the histograms which exhibits different bin values for two different frames taken from sample video.

3.2 Shot Boundary Detection

In this section, techniques to detect abrupt and gradual transitions has been carried out on subset videos of TRECVID 2001 and 2007 datasets. The task of abrupt transition detection has been accomplished using MPBW histograms obtained in the previous section and employing a distance comparison

Figure 6. Illustration of MPBW code value generation for frame #00645 of BG_3097 video sequence

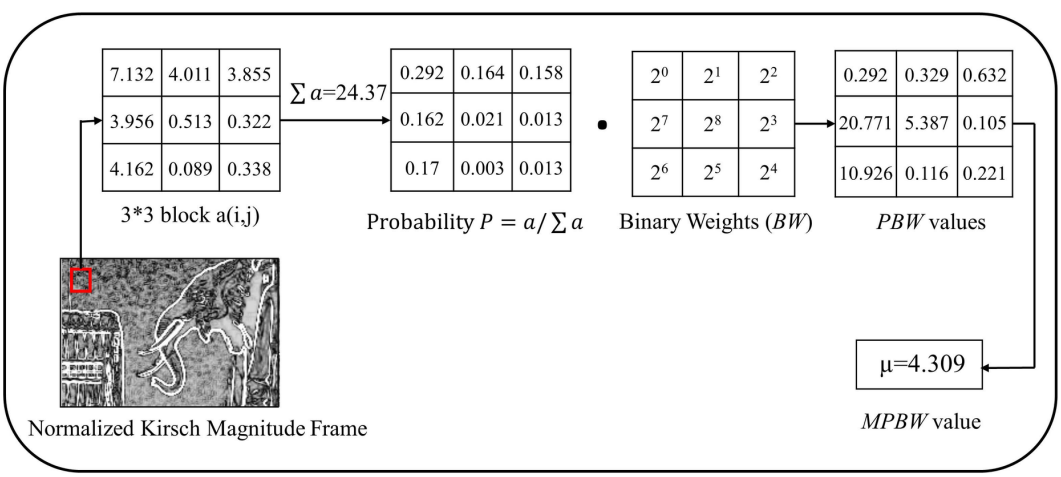
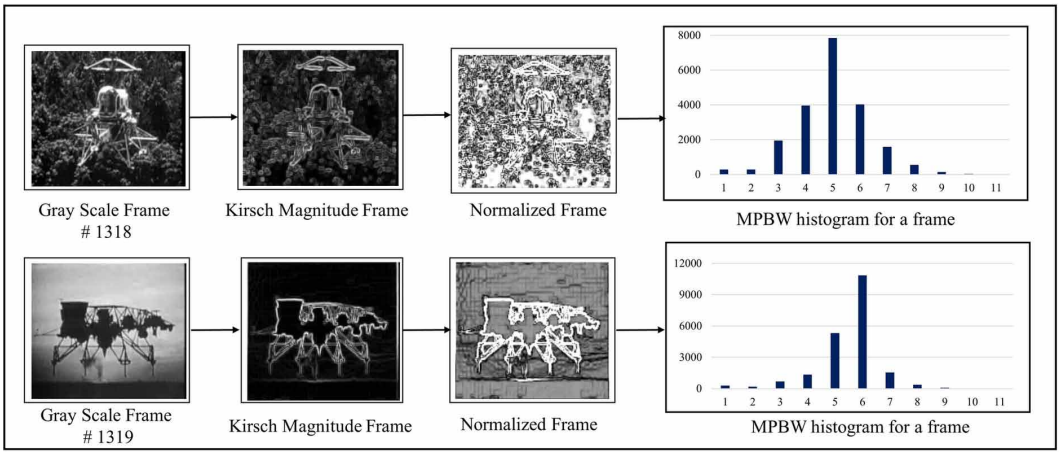


Figure 7. Illustration of MPBW Histogram for two different frames of anni006 video sequence



measure between the consecutive frames of the video. Whereas, gradual transitions are detected by computing the CMD and variance on every MPBW histograms. The proposed approach also provides a mechanism to detect and eliminate the unwanted frames in the video.

3.2.1 Identification and Elimination of Unwanted Frames

TRECVID datasets are challenging as it comprises object/camera motion and illumination variation along with the video editing effects. It is observed that, in some of the videos there is a possibility that the motion effect may lead to false transition detection and some of the blank frames acts as abrupt transition (Thounaojam et al., 2017). Thus, it is necessary to perform pre-processing operation to remove unwanted frames which leads to false transition prior to the abrupt and gradual transition detection in videos. This task is attained by applying variance statistical measure on the obtained MPBW histogram of every frame of a video.

Let H_k contain MPBW histogram values of a frame where $\{k=1, 2, 3,..., n\}$, then variance for each MPBW histogram is formulated as follows:

$$V_i = \frac{\sum_{k=1}^{n}\left(H_k - \mu\right)^2}{n} \tag{8}$$

where, $\mu = \frac{1}{n}\sum_{k=1}^{n} H_k$ and V_i is the variance of ith MPBW histogram in the video sequence.

Further, a threshold is set for every video sequence and employed to distinguish transition and non-transition frames by utilizing variance values for all MPBW histograms of a video and is formulated as follows:

$$TH_{NT} = (\mu + \sigma) * \alpha \tag{9}$$

where, TH_{NT} represents adaptive threshold value, μ is the mean value, α is a constant value and σ denotes Standard Deviation value of all variances computed for the video. If the non-transition frames are included for the shot boundary detection process, it will result in false transitions detection in the

videos. Margin of error varies from video to video depending on the number of non-transition frames present in that particular video. For example, Sample video BG_3314 of TRECVID 2007 benchmark dataset contains two blank frames in the beginning of the video sequence and fourteen blank frames at the end of the video sequence which doesn't belongs to either abrupt or gradual transition. Since blank frames differ from adjacent shot frames, it will be considered as a separate shot and algorithm detects the abrupt transitions. Thus, it is essential to identify and remove these non-transition frames prior detecting abrupt or gradual transitions.

3.2.2 Abrupt Transition Detection

The most significant way to find the quantitative change between adjacent frames is to compare their salient features. For abrupt shot detection, histogram metrics provides the best results, when computed at block level (Ford et al., 2000). In the proposed approach, the distance comparison between MPBW histogram features of the adjacent frames is computed to detect the abrupt cuts using Euclidean Distance which is formulated as follows:

$$DI_{(a,b)} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{10}$$

where *DI* is the distance measure and $\{a_i\},\{b_i\}$ are histogram values of adjacent frames of a video comprising of *n* frames.

Figure 8 shows the distance values between the consecutive frames of the sample video taken from TRECVID 2007 dataset where low and high peaks are clearly visible. High peaks represents an abrupt cut between two video shots whereas, low peaks depicts no transition within the shot.

Generally, threshold is used to determine transition between the consecutive frames of the video. Transitions are detected if the distance between the consecutive frames exceeds the set threshold. Therefore, it is very essential to select a proper threshold value to detect the abrupt/gradual transition accurately. It is a difficult task to set the hard threshold as the content and behaviour of the video varies from one another. Thus, an adaptive threshold is required to indicate the transitions.

In this abrupt transition approach, Mean (μ) and Standard Deviation (σ) of the distance values obtained along with a constant value 'α' are considered to compute the adaptive threshold as follows:

$$TH_{AT} = \left(\mu + \sigma\right) * \alpha \tag{11}$$

Here, the distance values that exceeds the set adaptive threshold TH_{AT} indicates abrupt transition in the video representing a camera break.

3.2.3 Gradual Transition Detection

In this section, an approach to detect gradual transitions in a video is discussed. Generally, gradual transition detection is more difficult when compared to abrupt transition due to object and camera motion (Hannane et al., 2016). This difficulty is due to the progressive change of the frame contents that has spread across the frames of the transition region (Bhaumik, Chakraborty, et al., 2016). Each category of gradual transition possess different patterns and it can be identified with its specific patterns (Rashmi & Nagendraswamy, 2020). During fade-in transition, the shot occurs gradually from a monochrome frame, usually a dark frame. Fade-out transition is just a reversal of fade-in where, the shot gradually disappears to a monochrome frame, usually a dark frame (Chavan & Akojwar, 2017). Dissolve transition is a combination of fade-in and fade-out where, the current shot gradually starts disappearing and the next shot gradually starts appearing simultaneously (Thounaojam et al., 2017). Figure 9 illustrates various gradual transition forms.

Figure 8. Distribution of distance values between the MPBW histograms of BG_8947

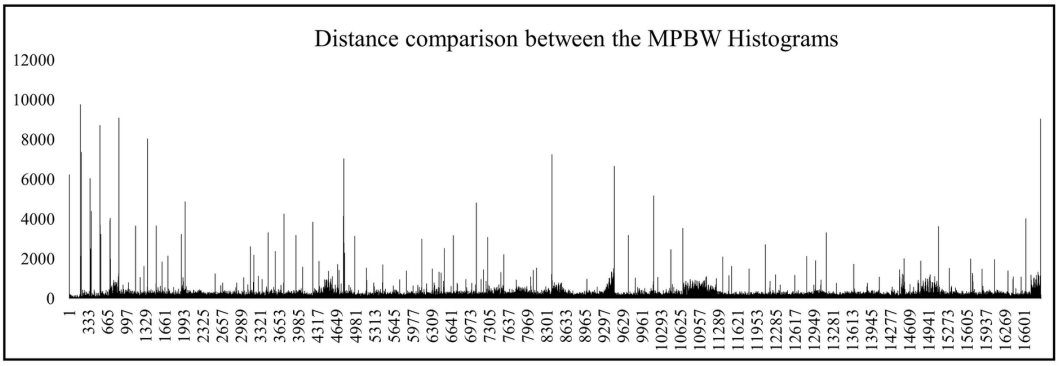
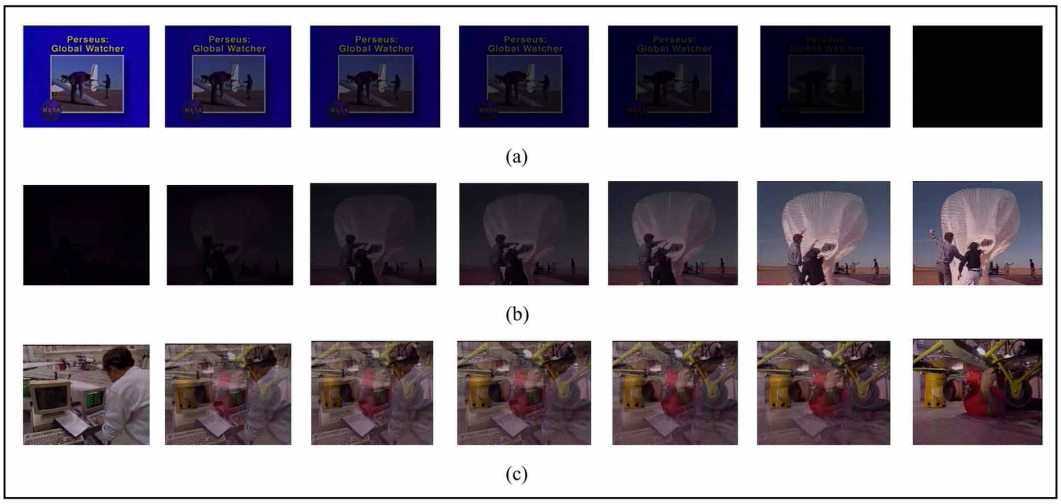


Figure 9. A set of consecutive frames extracted from nad58 video to illustrate (a) Fade-out (b) Fade-in and (c) Dissolve



Before implementing gradual transition algorithm, the abrupt transition and non-transition frames detected by utilizing adaptive threshold procedure as discussed in section 3.2.1 are excluded from the frame sequence.

Furthermore, a threshold is required to select the sequence of frames for the process of gradual transition detection. This task is accomplished using mean (μ) and Standard Deviation (σ) of the variance values of all MPBW histograms of a video and is formulated as follows:

$$TH_{GT} = \mu + \sigma \tag{12}$$

The frames that falls below the threshold TH_{GT} are considered for the process of gradual transition detection.

Gradual transition process usually takes place over multiple frames. Thus, to determine the gradual transition it is necessary to observe multiple frames for the specific pattern. The variance computed for MPBW histogram of each frame of a video is almost constant for the similar shot sequence whereas, it exhibits curve patterns for gradual transition. The fade-in and fade-out transition effects displays

linearly increasing and decreasing curve pattern respectively. Dissolve transition portrays near to parabolic characteristics which exhibits decreasing patterns in the beginning of frame sequence of changes and increasing patterns at appearance of new shot (Yoo et al., 2006).

Even though the variance feature of the MPBW histogram exhibits an ideal variance curve, an additional statistical measure is computed for the MPBW histogram values of the frame to reduce distortion within the curve and to obtain more accurate detection.

Detection of gradual transition is attained by computing variance by CMD values of each ith MPBW histogram and is formulated as follows:

$$HF_i = \frac{V_i}{CMD_i} \tag{13}$$

where, V_i *is* the variance values obtained in equation 8 of section 3.2.1. Further, CMD_i is the Co-efficient of Mean Deviation value which corresponds to the ith MPBW histogram and is formulated as follows:

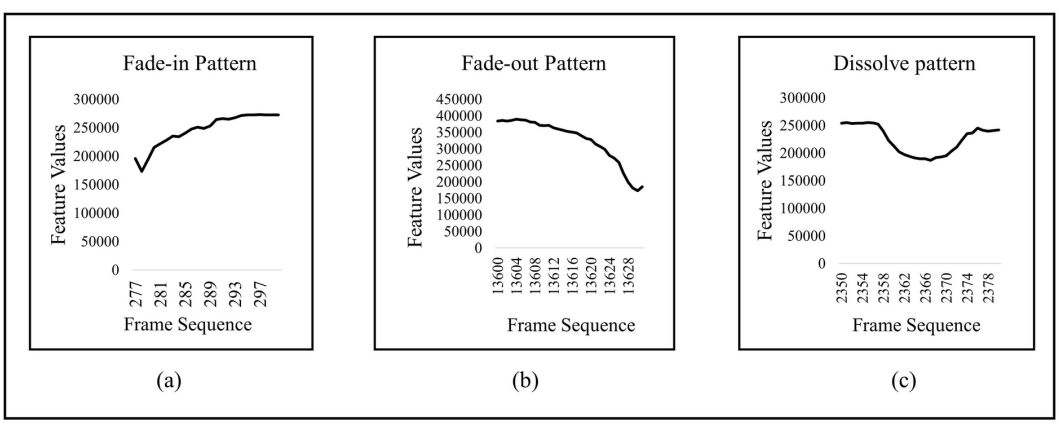
$$CMD_i = \frac{MD}{A} \tag{14}$$

where, A is the mean value and MD is the Mean Deviation value of each MPBW histogram and it is computed as follows:

$$MD = \frac{1}{n}\sum(H_k - A) \tag{15}$$

In the proposed method of gradual transition detection, HF_i represents the feature value for each frame in the sequence. The obtained HF_i values possess discriminating feature which helps to identify the inter-frame transition. The increasing or decreasing pattern represents different types of gradual transition unlike for a normal scene variation. The exclusive patterns are observed for consideration of fade-in, fade-out and dissolve transitions for sample frame sequence of NAd58 video are depicted in Figure10. The first increasing curve represents fade-in transition, decreasing curve is due to fade-

Figure 10. Illustration of (a) Fade-in (b) Fade-out and (c) Dissolve patterns from NAD58 video



out transition and last curve pattern is dissolve which is a combination of fade-in and fade-out (except blank frames). This step is repeated for all the remaining group of frames in the sequence and if, increasing or decreasing pattern is observed then, it can be considered as a gradual transition. The wipe gradual transition is not considered in the proposed approach.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Dataset: For experimentation and performance evaluation of the proposed method, video sequences from TRECVID 2001 and 2007 benchmark datasets have been considered. The description of the video datasets along with the number of abrupt cuts and gradual transitions is given in Table 1. TRECVID 2001 datasets are openly available and can be easily downloaded from "*Open Video Project*" whereas TRECVID 2007 datasets are obtained for research from *Netherlands Institute for Sound and Vision* under request.

Discussion: The quantitative evaluation metrics such as Recall, Precision and F1-score are used to evaluate the performance of the proposed approach which is computed as follows:

$$Recall = \frac{N_C}{N_C + N_M} \tag{16}$$

$$Precision = \frac{N_C}{N_C + N_F} \tag{17}$$

Table 1. Description of TRECVID 2001 and 2007 video datasets used for experimentation

Video	No. of Frames	Transition			Video Sources
		Abrupt	Gradual	Total	
Anni005	11362	38	27	65	TRECVID 2001
Anni006	16586	42	31	73	
Anni009	12304	39	64	103	
Anni010	31389	98	55	153	
NAD57	12510	45	26	71	
NAD58	13648	40	45	85	
BG_3027	49813	127	1	128	TRECVID 2007
BG_3097	44987	91	-	91	
BG_3314	35800	42	-	42	
BG_16336	2462	20	-	20	
BG_37309	9639	11	8	19	
BG_37770	15836	8	29	37	
BG_22677	15669	52	59	111	
BG_36658	26990	150	92	242	
BG_8947	16892	75	18	93	

$$F1-score = \frac{2 * \mathrm{Re}\,call * \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \tag{18}$$

where N_C is the number of correctly detected shot transitions, N_M is the number of missed shot transitions and N_F represents false transitions identified. *F1-score* is the harmonic mean of recall and precision that reflects on recall and precision rates. An outstanding shot transition detection approach must acquire high recall and high precision values. Extensive experiments were carried out to analyse the performance of the proposed approach using Matlab 2014a on Intel Core i5 processor, running at 2.70 GHz with 8GB RAM. The average algorithmic time complexity of the video depends on the number of frames to be processed in the particular video. The average time complexity increases linearly considering accountability of video frames. In the proposed method, algorithmic complexity depends on the video frame resolution pertaining to the dataset and in general, time complexity is of order $\theta(n^2)$.

4.1 Experimental Results on Abrupt Transition Detection

To detect abrupt transitions in video, an adaptive threshold mechanism is employed. An appropriate threshold value is set for each video by observing the peaks in the graph obtained for MPBW histogram distance values. To evaluate F1-score, threshold TH_{AT} is considered as described in section 3.2.2. The value of α lies in the range 0.1 to 2.5 for TRECVID 2001 dataset. Whereas, for TRECVID 2007 α value is in the range 0.1 to 1.5 and the experimental results are recorded. A comparative study is performed for the obtained results with some of the existing state-of-the-art SBD algorithms and are reported in Table 2 and 3 for TRECVID 2001 and 2007 datasets respectively. The results clearly exhibits that the proposed method outperforms all other methods during abrupt transition detection with 98.16% of average F1-score for TRECVID 2001 dataset and 96.32% of average F1-score for TRECVID 2007 dataset and it is graphically represented in Figure 11 and Figure 12 respectively.

4.2 Experimental Results on Gradual Transition Detection

In order to detect gradual transitions in video, an adaptive threshold is heuristically set based on mean and standard deviation of computed variance values of MPBW histograms of whole video. The frames which falls below the threshold are considered for the process of gradual transition detection.

In the proposed method, gradual transition patterns are generated for the feature values computed using variance and CMD values of MPBW histograms of frame sequence as formulated in the equation 12 of section 3.2.3. (Thounaojam et al., 2017) have noticed through experimentation that in TRECVID

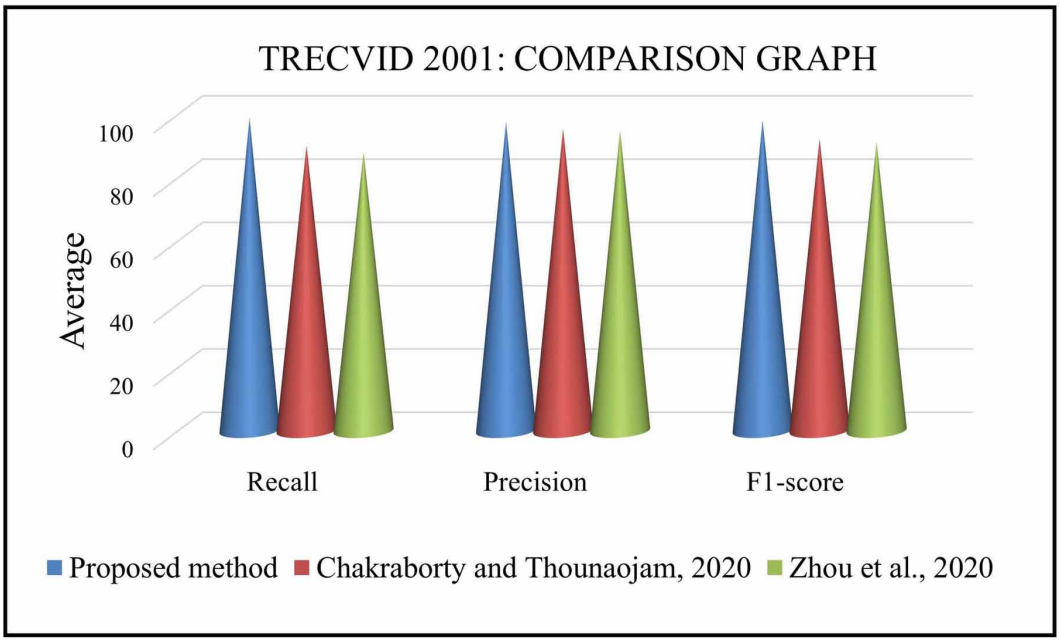
Table 2. Performance comparison with competing techniques for abrupt transition on TRECVID 2001 dataset

Video	Proposed Method			(Chakraborty & Thounaojam, 2020)			(Zhou et al., 2020)		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
ANNI005	100	97.4	98.6	90.5	76.0	82.6	85.0	85.0	85.0
ANNI006	97.5	95.2	96.3	76.2	100.0	86.5	-	-	-
ANNI009	100	97.4	98.6	89.5	97.1	93.2	84.6	94.3	89.2
ANNI010	95.9	97.9	96.8	89.8	100.0	94.6	-	-	-
NAD57	100	100	100	100.0	100.0	100.0	95.6	100	97.7
NAD58	100	97.5	98.7	95.0	100.0	97.4	86.5	100	92.8
Average	98.91	97.56	98.16	90.16	95.5	92.38	87.92	94.82	91.17

Table 3. Performance comparison with competing techniques for abrupt transition on TRECVID 2007 dataset

Video	Proposed Method			(Chakraborty & Thounaojam, 2020)			(Thounaojam et al., 2017)		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
BG_3027	93.7	96.7	95.1	87.3	100.0	93.2	94.5	90.2	92.3
BG_3097	91.2	97.6	94.3	87.9	100.0	93.6	86.8	98.7	92.4
BG_3314	85.7	97.2	91.1	83.3	100.0	90.9	78.6	94.3	85.7
BG_16336	95.0	100	97.4	90.0	100.0	94.7	95.0	100	97.4
BG_37309	100	100	100	100.0	100.0	100.0	100	84.6	91.7
BG_37770	100	100	100	100.0	100.0	100.0	100	88.9	94.1
Average	94.26	98.58	96.32	91.41	100.0	95.4	92.48	92.78	92.26

Figure 11. Comparison graph of proposed abrupt transition detection method with other methods on TRECVID 2001 dataset



video datasets, length of gradual transition normally ranges from 6 to 32 frames. Pertaining to this, in the proposed approach sample patterns were observed by grouping frames in the range of 5, 10, 15, 20, 25 and 32 at an instance of time. The task of gradual transition detection is attained by plotting variance by CMD values for the selected groups specifically. After thorough analysis, it is observed that our technique yields expected curve patterns for fade-in, fade-out and dissolve transitions in the range of 20-32 frames. Performance of the proposed approach is evaluated by determining F1-score and comparing it with some of the existing state-of-the-art gradual transition detection algorithms. The details of the results are reported in Table 4 and 5 for TRECVID 2001 and 2007 datasets respectively. Figure 13 and Figure 14 shows comparative results of our approach with regard to some of the existing approaches using TRECVID 2001 and 2007 datasets. It is evident that the proposed

Figure 12. Comparison graph of proposed abrupt transition detection method with other methods on TRECVID 2007 dataset

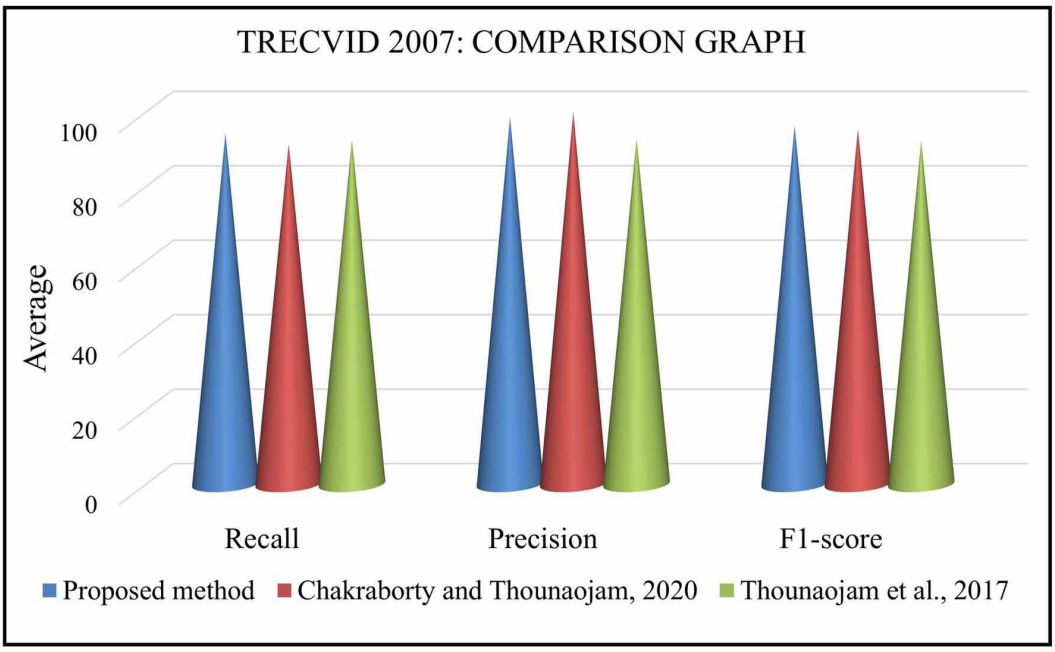


Table 4. Performance comparison with competing techniques for gradual transition on TRECVID 2001 dataset

Video	Proposed Method			(Chakraborty & Thounaojam, 2020)			(Zhou et al., 2020)		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
ANNI005	88.8	85.7	87.2	88.2	83.3	85.7	77.8	84.0	80.8
ANNI006	96.7	78.9	86.9	96.8	62.5	76.0	-	-	-
ANNI009	85.9	93.2	89.4	84.4	90.0	87.1	69.7	88.5	78.0
ANNI010	87.2	73.8	79.9	85.5	67.1	75.2	-	-	-
NAD57	92.3	85.7	88.8	84.6	78.6	81.5	84.0	72.4	77.8
NAD58	91.1	93.1	92.1	88.9	91.0	89.9	83.3	81.4	82.4
Average	90.3	85.06	87.38	88.06	78.75	82.56	78.7	81.57	79.75

gradual transition detection algorithm outperforms the compared approaches with 87.38% of average F1-score for TRECVID 2001 dataset and 82.9% of average F1-score for TRECVID 2007 dataset.

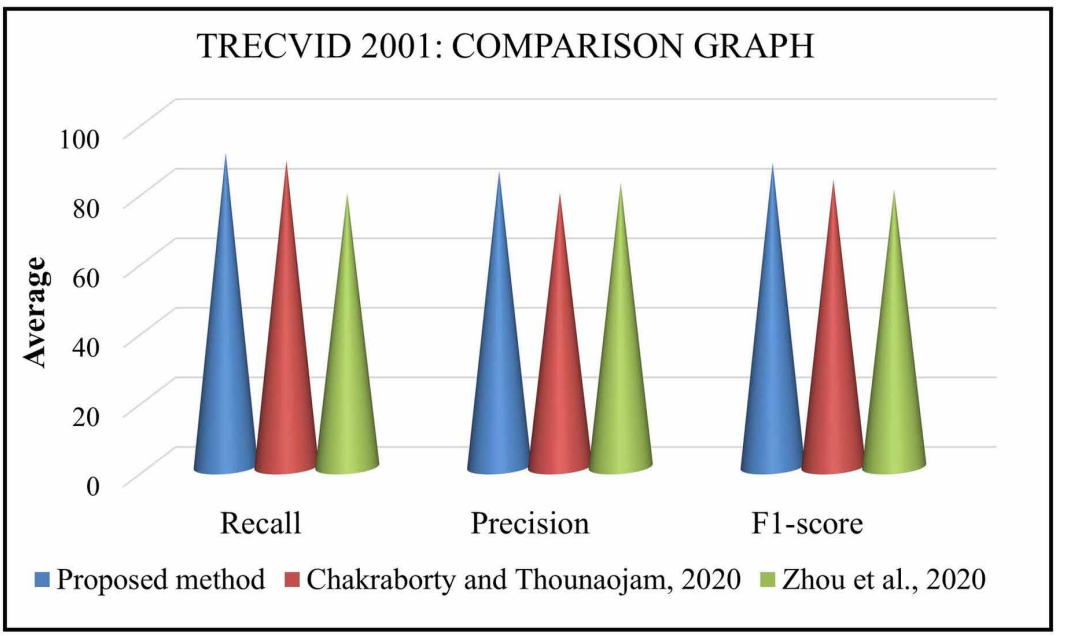
It can be summarised from the comparison tables that, the proposed method yields better results when compared with some of the existing SBD methods. Experimental results depicts high detection rates in terms of performance evaluation metrics in both abrupt and gradual transitions detection.

The proposed algorithm utilises both local and global features from frames to represent spatial resolution and thus acts less sensitive to object/camera motion. If frames of adjacent shots possess same histogram values there may be a chance of false transition detection. The proposed approach is computationally expensive when compared with global histogram techniques and the overall performance of algorithm is affected when operating with complex camera and light variations.

Table 5. Performance comparison with competing techniques for gradual transition on TRECVID 2007 dataset

Video	Proposed Method			(Chakraborty & Thounaojam, 2020)		
	Rec	Pre	F1	Rec	Pre	F1
BG_3027	100.0	50.0	66.7	100.0	50.0	66.7
BG_37309	87.5	70.0	77.8	87.5	63.6	73.7
BG_3770	93.1	84.3	88.4	92.0	79.3	85.2
BG_22677	86.4	89.4	87.9	85.4	88.3	86.8
BG_36658	88.0	89.0	88.5	85.8	87.7	86.7
BG_8947	83.3	93.7	88.1	72.2	100.0	83.9
Average	89.7	79.4	82.9	87.1	78.1	80.5

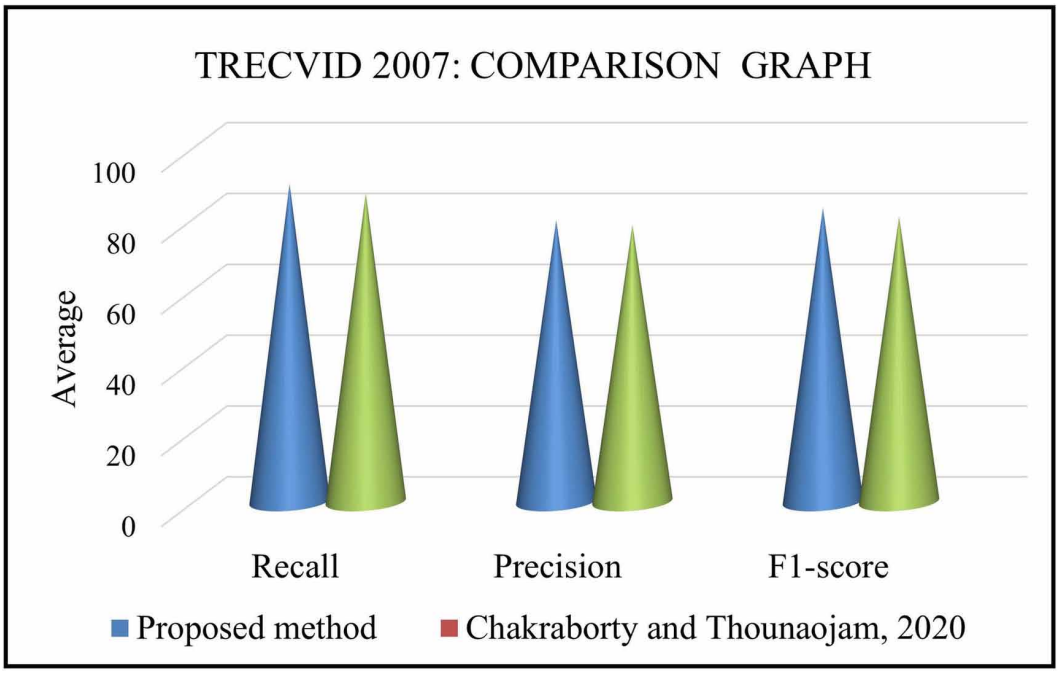
Figure 13. Comparison graph of proposed gradual transition detection method with other methods on TRECVID 2001 dataset



5. CONCLUSION

In this research work, an efficient approach to detect shot boundaries in videos is presented considering the combination of local and global features. The proposed method extracts block based MPBW histogram from each normalized Kirsch magnitude frame which has better discriminative capability. Abrupt cuts in video are identified by applying the distance measure between the MPBW histograms of the consecutive frames and employing adaptive threshold. Gradual transition patterns are detected by computing CMD and variance statistical measure on the MPBW histograms of the frame sequence and employing adaptive threshold. Experimental results depicts that the proposed method outperforms

Figure 14. Comparison graph of proposed gradual transition detection method with other methods on TRECVID 2007 dataset



some of the state-of-the-art approaches in terms of performance evaluation metrics using TRECVID 2001 and 2007 benchmark datasets. Future work includes the extension of the proposed work to address identification of wipe transitions in videos and reduction of computational complexity of the algorithm, so that the entire system can be applied to content based video analysis and retrieval systems.

ACKNOWLEDGMENT

Sound and Vision video is copyrighted. The Sound and Vision video used in this work is provided solely for research purposes through the TREC Video Information Retrieval Evaluation Project Collection.

REFERENCES

A., P., Mlsna, J., J., & Rodríguez. (2009). *Gradient and Laplacian Edge Detection*. Academic Press.

Adjeroh, D., Lee, M. C., Banda, N., & Kandaswamy, U. (2009). Adaptive edge-oriented shot boundary detection. *EURASIP Journal on Image and Video Processing*, *2009*, 1–13. Advance online publication. doi:10.1155/2009/859371

Bhaumik, H., Bhattacharyya, S., Das Nath, M., & Chakraborty, S. (2016). Hybrid soft computing approaches to content based video retrieval: A brief review. *Applied Soft Computing*, *46*, 1008–1029. doi:10.1016/j.asoc.2016.03.022

Bhaumik, H., Chakraborty, M., Bhattacharyya, S., & Chakraborty, S. (2016). Detection of gradual transition in videos: Approaches and applications. *Intelligent Analysis of Multimedia Information*. doi:10.4018/978-1-5225-0498-6.ch011

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *8*(6), 679–698. doi:10.1109/TPAMI.1986.4767851 PMID:21869365

Chakraborty, S., & Thounaojam, D. M. (2020). SBD-Duo: A dual stage shot boundary detection technique robust to motion and illumination effect. *Multimedia Tools and Applications*. Advance online publication. doi:10.1007/s11042-020-09683-y

Chakraborty, S., Thounaojam, D. M., & Sinha, N. (2020). A Shot boundary Detection Technique based on Visual Colour Information. *Multimedia Tools and Applications*. Advance online publication. doi:10.1007/s11042-020-09857-8

Chaudhary, M. D., & Upadhyay, A. B. (2014). Fusion of local and global features using stationary wavelet transform for efficient content based image retrieval. *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2014*. doi:10.1109/SCEECS.2014.6804471

Chavan, S., & Akojwar, S. (2017). *Effective Algorithm for Detection of Wipes in Presence of Motion and Illumination*. doi:10.2991/iccasp-16.2017.86

Dadashi, R., & Kanan, H. R. (2013). AVCD-FRA:A novel solution to automatic video cut detection using fuzzy-rule-based approach. *Computer Vision and Image Understanding*, *117*(7), 807–817. doi:10.1016/j.cviu.2013.03.002

Ford, R. M., Robson, C., Temple, D., & Gerlach, M. (2000). Metrics for shot boundary detection in digital video sequences. *Multimedia Systems*, *8*(1), 37–46. doi:10.1007/s005300050003

Guru, D. S., & Suhil, M. (2013). Histogram based split and merge framework for shot boundary detection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8284, 180–191. doi:10.1007/978-3-319-03844-5_19

Hanjalic, A. (2002). Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, *12*(2), 90–105. doi:10.1109/76.988656

Hannane, R., Elboushaki, A., Afdel, K., Naghabhushan, P., & Javed, M. (2016). An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram. *International Journal of Multimedia Information Retrieval*, *5*(2), 89–104. doi:10.1007/s13735-016-0095-6

Ishii, K., Sasaki, M., Matsui, M., Sakamoto, S., Yamaji, S., Hayashi, N., Mori, T., Kitagaki, H., Hirono, N., & Mori, E. (2000). A diagnostic method for suspected Alzheimer's disease using H2 15O positron emission tomography perfusion Z score. *Neuroradiology*, *42*(11), 787–794. doi:10.1007/s002340000404 PMID:11151682

Jadon, R. S., Chaudhury, S., & Biswas, K. K. (2001). A fuzzy theoretic approach for video segmentation using syntactic features. *Pattern Recognition Letters*, *22*(13), 1359–1369. doi:10.1016/S0167-8655(01)00041-1

Janwe, N. J., & Bhoyar, K. K. (2013). Video shot boundary detection based on JND color histogram. *2013 IEEE 2nd International Conference on Image Information Processing, IEEE ICIIP 2013*, 476–480. doi:10.1109/ICIIP.2013.6707637

Jiang, X., Sun, T., Liu, J., Chao, J., & Zhang, W. (2013). An adaptive video shot segmentation scheme based on dual-detection model. *Neurocomputing*, *116*, 102–111. doi:10.1016/j.neucom.2011.11.037

Kabbai, L., Abdellaoui, M., & Douik, A. (2018). Image classification by combining local and global features. *The Visual Computer*, *35*(5), 679–693. doi:10.1007/s00371-018-1503-0

Kanagaraj, K., & Priya, G. G. L. (2018). Curvelet transform based feature extraction and selection for multimedia event classification. *Journal of King Saud University - Computer and Information Sciences*. 10.1016/j.jksuci.2018.11.006

Kanan, C., & Cottrell, G. W. (2012). Color-to-grayscale: Does the method matter in image recognition? *PLoS One*, *7*(1), e29740. Advance online publication. doi:10.1371/journal.pone.0029740 PMID:22253768

Kirsch, R. A. (1971). Computer determination of the constituent structure of biological images. *Computers and Biomedical Research, an International Journal*, *4*(3), 315–328. doi:10.1016/0010-4809(71)90034-6 PMID:5562571

Muralidharan, R., & Chandrasekar, C. (2012). Combining local and global feature for object recognition using SVM-KNN. *International Conference on Pattern Recognition, Informatics and Medical Engineering, PRIME 2012*, 1–7. doi:10.1109/ICPRIME.2012.6208278

Nandini, H. M., Chethan, H. K., & Rashmi, B. S. (2020). Shot based keyframe extraction using edge-LBP approach. *Journal of King Saud University - Computer and Information Sciences*. 10.1016/j.jksuci.2020.10.031

Pal, G., Rudrapaul, D., Acharjee, S., Ray, R., Chakraborty, S., & Dey, N. (2015). Video shot boundary detection: A review. *Advances in Intelligent Systems and Computing*, *338*, 119–127. doi:10.1007/978-3-319-13731-5_14

Priya, G. G. L., & Domnic, S. (2012). Edge Strength Extraction using Orthogonal Vectors for Shot Boundary Detection. *Procedia Technology*, *6*, 247–254. doi:10.1016/j.protcy.2012.10.030

Rashmi, B. S., & Nagendraswamy, H. S. (2016). Abrupt shot detection in video using weighted edge information. *ACM International Conference Proceeding Series*. doi:10.1145/2980258.2980406

Rashmi, B. S., & Nagendraswamy, H. S. (2018). Effective Video Shot Boundary Detection and Keyframe Selection using Soft Computing Techniques. *International Journal of Computer Vision and Image Processing*, *8*(2), 27–48. doi:10.4018/IJCVIP.2018040102

Rashmi, B. S., & Nagendraswamy, H. S. (2020). Video shot boundary detection using block based cumulative approach. *Multimedia Tools and Applications*. Advance online publication. doi:10.1007/s11042-020-09697-6

Sasithradevi, A., & Mohamed Mansoor Roomi, S. (2020). A new pyramidal opponent color-shape model based video shot boundary detection. *Journal of Visual Communication and Image Representation*, *67*, 102754. Advance online publication. doi:10.1016/j.jvcir.2020.102754

Sengupta, A., Thounaojam, D. M., Singh, K. M., & Roy, S. (2015). Video shot boundary detection: A review. *Proceedings of 2015 IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2015*. doi:10.1109/ICECCT.2015.7226084

Shekar, B. H., & Uma, K. P. (2015). Kirsch Directional Derivatives Based Shot Boundary Detection: An Efficient and Accurate Method. *Procedia Computer Science*, *58*, 565–571. doi:10.1016/j.procs.2015.08.074

Singh, R. D., & Aggarwal, N. (2015). Novel research in the field of shot boundary detection – a survey. *Advances in Intelligent Systems and Computing*, *320*, 457–469. doi:10.1007/978-3-319-11218-3_41

Sintorn, I. M., Bischof, L., Jackway, P., Haggarty, S., & Buckley, M. (2010). Gradient based intensity normalization. *Journal of Microscopy*, *240*(3), 249–258. doi:10.1111/j.1365-2818.2010.03415.x PMID:21077885

Sobel, I., & Feldman, G. (1968). An Isotropic 3x3 Image Gradient Operator. *Stanford Artificial Intelligence Project*, (June), 271–272.

Thounaojam, D., Bhadouria, V. S., Roy, S., & Singh, K. M. (2017, June). Vivek, T., Bhadouria, S., Roy, S., & Singh, K. M. (2017). Shot boundary detection using perceptual and semantic information. *International Journal of Multimedia Information Retrieval*, *6*(2), 167–174. Advance online publication. doi:10.1007/s13735-017-0123-1

Thounaojam, D. M., Khelchandra, T., Singh, K. M., & Roy, S. (2016). A Genetic Algorithm and Fuzzy Logic Approach for Video Shot Boundary Detection. *Computational Intelligence and Neuroscience*, *2016*, 1–11. Advance online publication. doi:10.1155/2016/8469428 PMID:27127500

Thounaojam, D. M., Roy, S., Mohapatra, K. M. S. D. P., & Patnaik, S. (2014). A Survey on Video Segmentation. *Advances in Intelligent Systems and Computing*, *243*(December), V. doi:10.1007/978-81-322-1665-0

Venmathi, A. R., Ganesh, E. N., & Kumaratharan, N. (2016). Kirsch Compass Kernel Edge Detection Algorithm for Micro Calcification Clusters in Mammograms. *Middle East Journal of Scientific Research*, *24*(4), 1530–1535. doi:10.5829/idosi.mejsr.2016.24.04.23384

Yoo, H. W., Ryoo, H. J., & Jang, D. S. (2006). Gradual shot boundary detection using localized edge blocks. *Multimedia Tools and Applications*, *28*(3), 283–300. doi:10.1007/s11042-006-7715-8

Zhang, H., Wu, Q. M. J., & Nguyen, T. M. (2013). Incorporating mean template into finite mixture model for image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(2), 328–335. doi:10.1109/TNNLS.2012.2228227 PMID:24808286

Zhou, S., Wu, X., Qi, Y., Luo, S., & Xie, X. (2020). Video shot boundary detection based on multi-level features collaboration. *Signal, Image and Video Processing*. Advance online publication. doi:10.1007/s11760-020-01785-2

Nandini H. M. obtained her M.Sc degree from Amrita Vishwa Vidya Peetham University, Mysore, Karnataka, India in 2009. She is currently working as Assistant Professor in the department of Information Technology, Karnataka State Open University, Mysore, Karnataka, India. She is presently working towards her PhD degree in the research area of Computer Vision, Image/Video Processing and Pattern Recognition at Maharaja Research Foundation affiliated to University of Mysore.

Chethan H. K. completed B.Sc, M.Sc, and Ph.D. from the University of Mysore, Karnataka, India. Presently working as Professor at Maharaja Institute of Technology, Thandavapura, Karnataka, India. Guiding eight Ph.D. Students in several domains. Have successfully guided 4 Ph.D. students and several projects for bachelor's and master's students. He has published 66 papers in International conferences and Journals.

Rashmi B. S. obtained her PhD degree from University of Mysore, Mysore, Karnataka, India in 2020. She is currently working as Assistant Professor in the department of Information Technology, Karnataka State Open University, Mysore, Karnataka, India. Her focused are of research is Content Based Image/Video Retrieval Systems, Pattern Recognition, Texture Analysis and Fuzzy Theory.

Smart Innovation, Systems and Technologies 251

Tomonobu Senjyu
Parakshit Mahalle
Thinagaran Perumal
Amit Joshi *Editors*

IOT with Smart Systems

Proceedings of ICTIS 2021, Volume 2

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 251

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at https://link.springer.com/bookseries/8767

Tomonobu Senjyu · Parakshit Mahalle ·
Thinagaran Perumal · Amit Joshi
Editors

IOT with Smart Systems

Proceedings of ICTIS 2021, Volume 2

Springer

Editors
Tomonobu Senjyu
University of the Ryukyus
Nishihara, Okinawa, Japan

Parakshit Mahalle
Sinhgad Technical Education Society
SKNCOE, Pune, India

Thinagaran Perumal
University Putra Malaysia Serdang
Serdang, Malaysia

Amit Joshi
Global Knowledge Research Foundation
Ahmedabad, India

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-981-16-3944-9 ISBN 978-981-16-3945-6 (eBook)
https://doi.org/10.1007/978-981-16-3945-6

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The Fifth International Conference on Information and Communication Technology for Intelligent Systems (ICTIS 2021) targets state-of-the-art as well as emerging topics pertaining to information and communication technologies (ICTs) and effective strategies for its implementation for engineering and intelligent applications.

The conference is anticipated to attract a large number of high-quality submissions, stimulate the cutting-edge research discussions among many academic pioneering researchers, scientists, industrial engineers, students from all around the world and provide a forum to researcher; propose new technologies, share their experiences and discuss future solutions for design infrastructure for ICT; provide a common platform for academic pioneering researchers, scientists, engineers and students to share their views and achievements; enrich technocrats and academicians by presenting their innovative and constructive ideas; focus on innovative issues at international level by bringing together the experts from different countries.

The conference was held during April 23–24, 2021, digitally on Zoom and organized by Global Knowledge Research Foundation in collaboration with KCCI and IFIP INTERYIT.

Research submissions in various advanced technology areas were received, and after a rigorous peer-review process with the help of program committee members and external reviewer, 160 papers were accepted with an acceptance rate of 16%. All 160 papers of the conference are accommodated in 2 volumes, and also papers in the book comprise authors from 18 countries.

This event success was possible only with the help and support of our team and organizations. With immense pleasure and honor, we would like to express our sincere thanks to the authors for their remarkable contributions, all the Technical Program Committee members for their time and expertise in reviewing the papers within a very tight schedule and the publisher Springer for their professional help.

We are overwhelmed by our distinguished scholars and appreciate them for accepting our invitation to join us through the virtual platform and deliver keynote speeches and technical session chairs for analyzing the research work presented by the researchers. Most importantly, we are also grateful to our local support team for

their hard work for the conference. This series has already been made a continuous series which will be hosted at different location every year.

Nishihara, Japan	Tomonobu Senjyu
Pune, India	Parakshit Mahalle
Serdang, Malaysia	Thinagaran Perumal
Ahmedabad, India	Amit Joshi
April 23–24, 2021	

Contents

About the Editors

Dr. Tomonobu Senjyu received his B.S. and M.S. degrees in Electrical Engineering from the University of the Ryukyus in 1986 and 1988, respectively, and his Ph.D. degree in Electrical Engineering from Nagoya University in 1994. Since 1988, he has been with the Department of Electrical and Electronics Engineering, University of the Ryukyus, where he is currently Professor. His research interests include stability of AC machines, power system optimization and operation, advanced control of electrical machines and power electronics. He is Member of the Institute of Electrical Engineers of Japan and IEEE.

Dr. Parakshit Mahalle holds a B.E. degree in CSE and an M.E. degree in Computer Engineering. He completed his Ph.D. at Aalborg University, Denmark. Currently, he is working as Professor and Head of the Department of Computer Engineering at STES Smt. Kashibai Navale College of Engineering, Pune, India. He has over 18 years of teaching and research experience. Dr. Mahalle has published over 140 research articles and eight books and has edited three books. He received the "Best Faculty Award" from STES and Cognizant Technologies Solutions.

Dr. Thinagaran Perumal received his B.Eng., M.Sc. and Ph.D. Smart Technologies and Robotics from Universiti Putra Malaysia in 2003, 2006 and 2011, respectively. Currently, he is Associate Professor at Universiti Putra Malaysia. He is also Chairman of the TC16 IoT and Application WG National Standard Committee and Chair of IEEE Consumer Electronics Society Malaysia Chapter. Dr. Thinagaran Perumal is Recipient of 2014 IEEE Early Career Award from IEEE Consumer Electronics Society. His recent research activities include proactive architecture for IoT systems, development of the cognitive IoT frameworks for smart homes and wearable devices for rehabilitation purposes.

Dr. Amit Joshi is currently Director of the Global Knowledge Research Foundation. He is also Entrepreneur and Researcher, and he holds B.Tech., M.Tech. and Ph.D. degrees. His current research focuses on cloud computing and cryptography. He is Active Member of ACM, IEEE, CSI, AMIE, IACSIT, Singapore, IDES, ACEEE, NPA and several other professional societies. He is also International Chair

of InterYIT at the International Federation of Information Processing (IFIP, Austria). He has published more than 50 research papers, edited 40 books and organized over 40 national and international conferences and workshops through ACM, Springer and IEEE across various countries including India, Thailand, Egypt and Europe.

Chapter 28
Video Shot Retrieval Using Multi-feature Approach

H. M. Nandini, H. K. Chethan, and B. S. Rashmi

Abstract The evolution of Internet and mobile technologies has raised the growth of video data that spur demand for efficient browsing and retrieval technologies. Content-Based Video Retrieval (CBVR) is evolved progressively to retrieve desired videos proficiently from large repositories based on the video content. In this aspect, an efficient CBVR method is presented using multiple features from video shot keyframes. The approach constructs histogram for Sobel magnitude of V-channel from HSV image and also local binary cumulative sum variance pattern for grayscale image. Histograms, thus, constructed are concatenated to build multiple feature vector database. Further, shot matching process is established by applying Euclidean distance between shot keyframe and query keyframe features. Experimentation was carried out on UCF YouTube action benchmark dataset to analyze the efficiency of the proposed algorithm. Video shot retrieved results depict significant improvement of the presented CBVR approach in comparison with baseline algorithm in terms of evaluation metric.

28.1 Introduction

In recent years, advancement in technologies and usage of Internet have led to extensive growth of video data on storage platforms. Consequently, such huge video repositories must be accompanied by efficient video retrieval and analysis technologies. Traditional search schemes are based on textual information of videos and show limited retrieval capability when they lack to understand the video content [1]. Thus, CBVR approaches have attained vital consideration in research field which exploits video content for the purpose of retrieval. CBVR systems have an extensive scope of implementation like digital museums, news event exploration [2], computer-aided retinal surgery [3], and CCTV surveillance systems [4].

H. M. Nandini (✉) · H. K. Chethan
Maharaja Research Foundation, Mysuru, India

H. M. Nandini · B. S. Rashmi
Karnataka State Open University, Mysuru, India

T. Senjyu et al. (eds.), *IOT with Smart Systems*, Smart Innovation,
Systems and Technologies 251, https://doi.org/10.1007/978-981-16-3945-6_29

The content in videos is broadly categorized into visual, textual, and audio contents. Out of these, visual content provides abundant information, and it is most frequently chosen for video retrieval [5]. There are numerous ways to extract visual content/feature of images such as color, texture, shape, and edges. The image/frame content is efficiently represented using edge feature and is utilized in various research works [6, 7]. Contents of image can be obtained by extracting features either globally or locally [8]. Global feature represents the visual information of the whole image in a single vector and is sensitive to noise, illumination, variation, etc. Their drawbacks are fixed by the use of local features that describe visual content in patches and encode the local information to get the finest details of the image [9]. Hence, several existing approaches have amalgamated both local and global features to obtain better results in various domains like retrieval of images [10] and detection of objects [11]. Use of multiple features enhances the accuracy of CBVR in terms of performance evaluation metric compared to classical method that utilizes color information only [12]. Motivated from these advantages, authors have carried out the proposed CBVR approach.

Generally, features are extracted from keyframes, as it contains most of the salient information of a video shot. This concept decreases computational complexity and preserves memory as well [13]. Thus, in this proposed approach, keyframes are selected from shot-based method proposed by Nandini et al. [7] to extract features for retrieval purpose. The proposed retrieval method extracts multiple features by incorporating both global and local features of shot keyframes. Two different approaches are explored for feature extraction: In the first method, initially, RGB images are transformed into HSV color images. Further, Sobel gradient magnitude for V-channel of HSV color images is extracted to construct the histograms. In the second method, RGB images are transformed to gray images, and Local Binary Cumulative Sum Variance (LBCSV) patterns are extracted to construct the histogram. These two histogram features are concatenated to construct the feature vector database. Subsequently, multi-feature vectors are used to match the query keyframe and shot keyframes by measuring Euclidean distance. The efficacy of the presented algorithm is verified using evaluation metrics by conducting experiments on UCF YouTube action benchmark datasets.

The article is organized as follows: Sect. 28.2 reports overview of some existing works. Section 28.3 provides proposed methodology covering description of feature extraction and similarity measurement. Section 28.4 presents experimental results and discussion, followed by conclusion of the paper in Sect. 28.5.

28.2 Related Works

An overview of existing works on CBVR approaches by various researchers is presented in this section. The importance and popularity of CBVR systems have led to several survey papers [14, 15]. An efficient CBVR approach has been presented [16] by extracting SURF features and adding color information using C-SURF of

every keyframes. Distance measure is applied to match the features between stored video data and test video to retrieve the most similar videos based on the rank. Rashmi et al. [17] proposed video retrieval approach by utilizing visual features of keyframes. Here, feature matrix is constructed by extracting color and edge features of all keyframes of the shots to retrieve similar videos. FALKON [18] is an approach for efficient CBVR system that uses deep features and distributed in-memory computing along with the power of big data technologies. Further, video query maps are introduced to enhance the efficiency of the retrieval system. An approach for effective video retrieval has been introduced [19] by utilizing spatiotemporal features. In this method, combination of motion feature vector and HSV color histogram is used to generate hybrid feature vector that plays a significant role in CBVR. Padmakala [20] has used multiple features of every frame and twelve distance measures that are optimized by gravitational search algorithm to extract relevant videos from the database based on the query. An efficient CBVR method is proposed [21] by extracting multiple features like motion, color, and texture features and employing similarity measure using Euclidean distance between query and video database. It can be observed from the above overview that utilization of multiple feature plays significant role for effective CBVR during the last few years.

28.3 Proposed Methodology

In this section, an efficient CBVR approach using multiple features of video shot keyframes has been presented. The fundamental stages of proposed CBVR method are feature extraction and similarity measurement. A schematic representation of the presented approach is illustrated in Fig. 28.1.

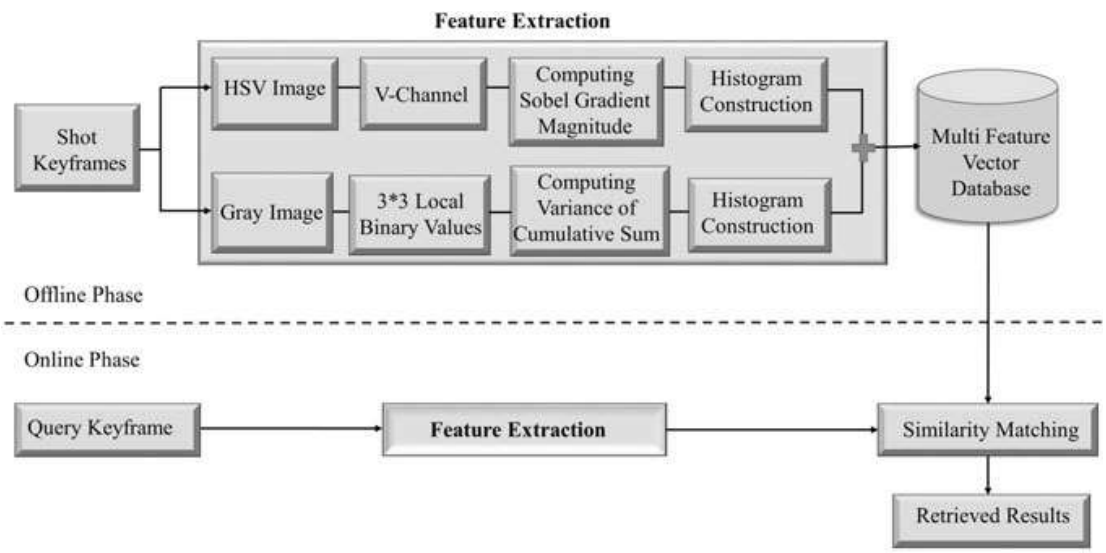


Fig. 28.1 Structure of the presented CBVR approach

28.3.1 Multi-feature Extraction

The visual features of images play a vital role in CBVR. The following subsections describe the process of multiple feature extraction from keyframes to achieve shot retrieval task.

Histogram Construction Using HSV Color Edge Approach
Edge detection helps to remove the irrelevant information and retains essential structural properties of the image [22]. Gray scale frame contains less information than color frame; thus, color edge detection is used to get more precise information [23]. There exists various types of color spaces in literature with respect to different applications. However, HSV color space has huge influence as it arranges any color image in the similar way that human eyes can perceive [8]. In the proposed method, V-channel of HSV color space is considered for edge detection, as it represents luminance of an image and stores the brightness intensity values. Sobel edge detector is employed on the V-channel of every keyframe to get Sobel magnitude as it performs better than other edge descriptors with reference to computational complexity and accuracy [24]. It also has an advantage of smoothing effect to the random noises present in the image [25]. The Sobel detector includes a two 3 * 3 kernels which are convolved with V-channel of the frame to measure derivatives approximations. Kernels produce two different measurements in vertical and horizontal directions for gradients, specifically G_x and G_y and are computed as:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad G_y = \begin{bmatrix} +1 & 2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{28.1}$$

The gradient magnitude provides rate of change for gradient in frame intensity at every pixel position as illustrated in Fig. 28.2 and is computed using following equation:

$$\mathrm{GM} = \sqrt{G_x^2 + G_y^2} \tag{28.2}$$

Histogram for every shot keyframe is built using gradient magnitude values.

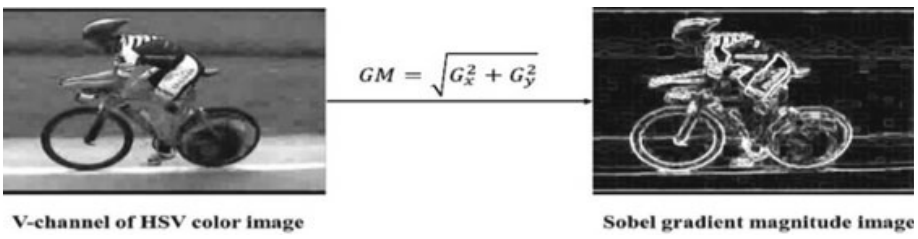


Fig. 28.2 Illustration of Sobel gradient magnitude for frame # 21_7_008 of biking video sequence

Histogram Construction Using LBCSV Pattern

The initial step of this feature extraction process starts by transforming shot keyframes from RGB to gray scale. Further, algorithm as applied in [26] is used to compute local binary values of frame by thresholding the 3 * 3 block of every pixel with the center pixel, and it is formulated as follows:

$$\mathrm{LBV}_{(P\,R)} = S(G_C - G_P) \tag{28.3}$$

$$s(a) = \begin{Bmatrix} 1 \; a \geq 0 \\ 0 \; a < 0 \end{Bmatrix}$$

where P and R represent number of pixels and radius, respectively. G_P is neighboring pixel values, and G_C represents center pixel value of 3 * 3 neighborhood. In this work, instead of computing LBP code on 3 * 3 local binary values, cumulative sum is evaluated using following equation:

$$\mathrm{CS}(n) = \sum_{p=1}^{n} (\mathrm{LBV}) \tag{28.4}$$

where LBV is binary values of each block and CS(n) represents cumulative sum. Furthermore, block variance of CS(n) is formulated as follows:

$$V = \frac{\sum_{k=1}^{n} (\mathrm{CS} - \mu)^2}{n} \tag{28.5}$$

where $\mu = \frac{1}{n} \sum_{k=1}^{n} \mathrm{CS}(i)$ and $n = 9$. This process is repeated at every pixel position in an overlapping fashion for the entire image. Finally, histogram of Local Binary Cumulative Sum Variance (LBCSV) values for each keyframe is constructed and is shown in Fig. 28.3.

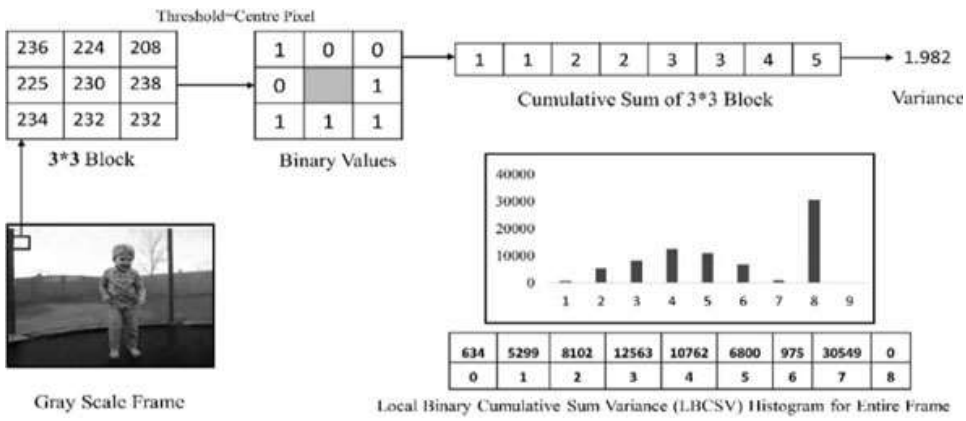


Fig. 28.3 Construction of LBCSV histogram for frame # 7_1_75 of trampoline jumping video

Histograms obtained from the above two feature representation methods are concatenated to build a multi-feature vector and are stored in database.

28.3.2 Similarity Measurement

Video shot retrieval result mainly confides in similarity measures. The similarity can be attained by matching objects, texts, features, etc., and combinations of them. However, the most direct and convenient method for measuring similarity is by matching features [15]. Extensive analysis of video retrieval similarity measures has revealed that Euclidean and Manhattan distance measures are the best in terms of their retrieval ability [27]. In the proposed approach, similarity between shot keyframe features in database and query keyframe feature is measured by Euclidean distance using following equation:

$$\mathrm{ED} = \sqrt{\sum_{i=1}^{n} (V_d - V_q)^2} \tag{28.6}$$

where, ED is the distance measure and V_{d} and V_{q} represent the multi-feature vectors. If the Euclidean distance is smaller, then database shot keyframe which corresponds to video is more similar to query keyframe and is higher in rank [28]. This concept is applied, and ranking of retrieved results is considered.

28.4 Results and Discussion

Dataset: The performance evaluation of the proposed retrieval approach is carried out considering video sequences from UCF YouTube action dataset and is described in Table 28.1. It contains 11 action categories and 1597 videos in total with 320 × 240 frame dimension.

28.4.1 Performance Evaluation

In order to check the efficiency of the presented algorithm, the query shot keyframe is randomly selected from the standard dataset. The multi-feature vector of query keyframe is matched with features in repository using similarity measurement as discussed in Sect. 28.2. Based on matching, the most similar video shots are retrieved and are ranked for analysis purpose. Performance evaluation of the presented method is determined by average precision score and compared with the existing baseline approach [29] considering top ten ranking results and is recorded in Table 28.2. The performance is graphically represented in Fig. 28.4. The result analysis based on the number of keyframes between query and repository is considered as in [29].

Table 28.1 Description of UCF YouTube action video dataset

Categories of video	No. of videos
“Biking”	137
“Diving”	145
“Golf_ swing”	156
“Soccer_ juggling”	142
“Trampoline_ jumping”	198
“Horse _riding”	156
“Basketball”	138
“Volleyball_ spiking”	167
“Swing”	119
“Tennis _swing”	116
“”Walking”	123

Table 28.2 Performance analysis using average precision score

Categories of video	Mallick et al. [29]	Proposed method
“Biking”	94.09	95.74
“Diving”	96.92	97.48
“Golf_ swing”	89.52	91.67
“Soccer_ juggling”	95.97	95.55
“Trampoline_ jumping”	97.74	98.34
“Horse _riding”	89.31	90.13
“Basketball”	96.43	98.11
“Volleyball_ spiking”	97.12	98.09
“Swing”	98.21	98.95
“Tennis _swing”	93.21	94.51
“”Walking”	96.32	97.23
Average	94.99	95.98

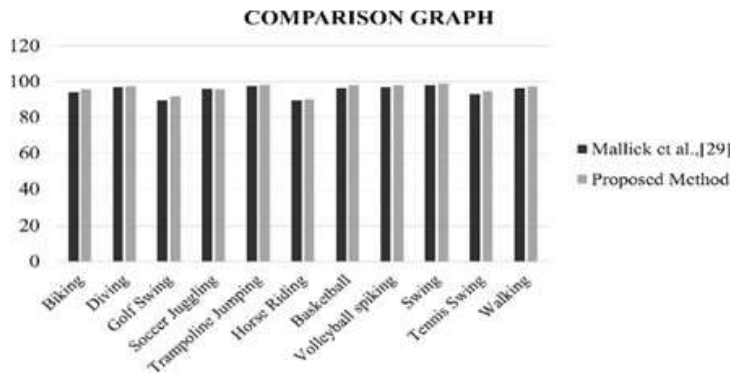


Fig. 28.4 Comparison graph of average precision score

In this approach, computation of multi-feature vector is much faster with feature dimension of 1 * 19 for each shot keyframe when compared to state-of-the-art algorithm [29] with 1 * 64 dimension. The results clearly exhibit that the proposed approach outperforms [29] with 95.98% of average precision score. The proposed approach is computationally expensive when compared with traditional system that uses single feature only.

28.5 Conclusion

In the proposed research work, video shots are retrieved utilizing visual content of shot keyframes. The method extracts multiple features by exploiting Sobel gradient magnitude for V-channel of HSV color images and extracting Local Binary Cumulative Sum Variance (LBCSV) patterns from gray images to construct the histograms. Similarity matching is employed by applying Euclidean distance between query keyframe and shot keyframe feature vectors present in the database to retrieve the similar video shots of the highest ranking. Experimental results show that the presented algorithm provides better result in comparison with baseline algorithm with 95.98% average precision score using UCF YouTube action benchmark dataset. The main drawback of this approach lies in multi-feature extraction which makes algorithm computationally complex. In the future work, the performance analysis of the presented algorithm will be investigated considering motion feature of the video sequence.

References

1. Rajendran, P.: An Enhanced Content-Based Video Retrieval System Based on Query Clip, vol. 1, pp. 236–253 (2009)
2. Peng, Y., Ngo, C.W.: Hot event detection and summarization by graph modeling and matching. Lect. Notes Comput. Sci. **3568**, 257–266 (2005)
3. Quellec, G., Lamard, M., Cazuguel, G., Droueche, Z., Roux, C., Cochener, B.: Real-time retrieval of similar videos with application to computer-aided retinal surgery. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4465–4468 (2011)
4. Yang, Y., Lovell, B.C., Dadgostar, F.: Content-based video retrieval (CBVR) system for CCTV surveillance videos. In: DICTA 2009—Digital Image Computing: Techniques and Applications, pp. 183–187 (2009)
5. Yasin, D., Sohail, A., Siddiqi, I.: Semantic video retrieval using deep learning techniques. In: Proceedings 2020 17th International Bhurban Conference on Applied Science and Applied Technologies (IBCAST 2020), pp. 338–343 (2020)
6. Patel, B.V., Deorankar, A.V., Meshram, B.B.: Content based video retrieval using entropy, edge detection, black and white color features. In: ICCET 2010, vol. 6, pp. 272–276 (2010)
7. Nandini, H.M., Chethan, H.K., Rashmi, B.S.: Shot based keyframe extraction using edge-LBP approach. J. King Saud Univ. Comput. Inf. Sci. (2020)

8. Thounaojam, D.M., Khelchandra, T., Singh, K.M., Roy, S.: A genetic algorithm and fuzzy logic approach for video shot boundary detection. Comput. Intell. Neurosci. **2016** (2016)
9. Kabbai, L., Abdellaoui, M., Douik, A.: Image classification by combining local and global features. Vis. Comput. **35**, 679–693 (2018)
10. Chaudhary, M.D., Upadhyay, A.B.: Fusion of local and global features using stationary wavelet transform for efficient content based image retrieval. In: 2014 IEEE Students' CEECS 2014 (2014)
11. Muralidharan, R., Chandrasekar, C.: Combining local and global feature for object recognition using SVM-KNN. Int. Conf. Pattern Recogn. Inf. Med. Eng. PRIME **2012**, 1–7 (2012)
12. Zaini, H.G., Frag, T.: Multi feature content based video retrieval using high level semantic concepts. Int. J. Comput. Sci. Issues **2**, 254–260 (2014)
13. Prathiba, T., Kumari, R.S.S.P.: Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer interaction. J. Ambient Intell. Humaniz. Comput. (2020).
14. Spolaôr, N., Diana, H., Shoity, W., Takaki, R., Augusto, L., Saddy, C., Coy, R., Chung, F.: Engineering Applications of Artificial Intelligence A Systematic Review on Content-Based Video Retrieval, vol. 90 (2020)
15. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **41**, 797–819 (2011)
16. Patel, D.H.: Content based video retrieval using enhance feature extraction. Int. J. Comput. Appl. **119**, 4–8 (2015)
17. Rashmi, M., Fernandes, R.: Video Retrieval Using Fusion of Visual Features and Latent Semantic Indexing, pp. 4272–4277 (2014)
18. Khan, M.N., Alam, A., Lee, Y.K.: FALKON: large-scale content-based video retrieval utilizing deep-features and distributed in-memory computing. In: Proceedings—2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 36–43 (2020)
19. Naveen Kumar, G.S., Reddy, V.S.K.: A scheme for shot detection and video retreival using spatio temporal features. Int. J. Recent Technol. Eng. **8**, 605–608 (2019)
20. Padmakala, S., Mala, G.S.A., Anandkumar, K.M.: Using diverse set of features to design a content-based video retrieval system optimised by gravitational search algorithm. Int. J. Bus. Intell. Data Min. **17**, 444–470 (2020)
21. Asha, D., Latha, Y.M., Reddy, V.S.K.: Content based video retrieval system using multiple features. Int. J. Pure Appl. Math. **118**:287–294 (2018)
22. Dhagdi, S.T., Deshmukh, P.R.: Keyframe based video summarization using automatic threshold & edge matching rate. Int. J. Sci. Res. Publ. **2**, 1–12 (2012)
23. Koschan, A.: A comparative study on color edge detection. In: Proceedings 2nd Asian Conference on Computer Vision, vol. 3, pp. 574–578 (1995)
24. Abdesselam, A.: Improving local binary patterns techniques by using edge information. Lect. Notes Softw. Eng., 360–363 (2013)
25. Rami, H., Hamri, M.: Color edge detector with Sobel-PCA. Int. J. Comput. Appl. **75**, 12–16 (2013)
26. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern. Recognit. **29**, 51–59 (1996)
27. Bekhet, S., Ahmed, A.: Evaluation of similarity measures for video retrieval. Multimed. Tools Appl. **79**, 6265–6278 (2020)
28. Patel, B.V.: Content based video retrieval systems. Int. J. UbiComp. **3**, 13–30 (2012)
29. Mallick, A.K., Mukhopadhyay, S.: Video retrieval using salient foreground region of motion vector based extracted keyframes and spatial pyramid matching. Multimed. Tools Appl. **79**, 27995–28022 (2020)