IMPROVING THE EFFICIENCY OF KEY FRAME EXTRACTION USING HYBRID MOTION VECTORS

Darshankumar D.Billur[1] and Dr.Manu T.M.[2]

[1]Asst.Professor, KLE College of Engineering & Technology, Chikodi
[2] Prof, & HOD ECE, KLE Institute of Technology, Hubli

ABSTRACT

Extracting relevant points of action from video sequences has found its applications in both commercial and non-commercial scenarios. Applications like CCTV monitoring, video summarization, video codec optimization, etc. make use of key frame extraction (KFE) to function effectively. KFE has also found its applicability in military applications where military training footages are given for KFE and the output is used for fast track training of the officers. In this paper, we propose a novel hybrid motion vector based KFE algorithm, that utilizes motion vector information and combines it with a multi-color space visual attention model to extract key frames. The proposed algorithm can improve the precision, recall and f-measure values when compared with state-of-the-art algorithms like Delaunay clustering, VSUMM, DT, OV. It is found that the proposed algorithm improves the efficiency of KFE by more than 20% across different video datasets.

KEYWORDS

Key frame extraction, summarization, multi-color space, motion vector

1. INTRODUCTION

Key frame extraction or KFE is the process where a video sequence is divided into frames, then these frames are processed one-by-one in order to find out redundant or not-so-useful frames. The removed frames are called non-key frames, while the retained frames are called as key frames. This process can be depicted with the help of the following diagram,

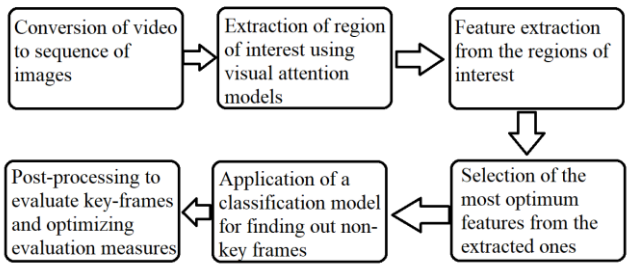

Figure 1. Process of key-frame extraction

The input frames are extracted from an input video sequence and are given to a region of interest extraction unit. This unit uses the concepts like visual attention models, motion vectors, color spaces, thresholding, edge mapping, and others to find out the areas of interest in the given frame. These areas of interest can contain objects, textures, important markers in the image, and other visually distinct areas.

K.S.Babai,P.K.Suresh,K.UmaRani et al. (Eds): ICARTEC-2021
pp. 01 - 09, 2021.IJCI – 2021 DOI:10.5121/ijci.2021.100301

Feature extraction methods like color maps, edge maps, gray level co-occurrence matrices, principal component analysis, and others are applied to the extracted regions. These methods perform the task of describing the image in numerical form. This numerical data can then be used for further processing. Due to application of a large number of feature extraction algorithms on the image, the size of the feature vector increases exponentially. Due to this, there is a need for feature selection.

Methods like variance evaluation, standard deviation calculation, and others are used for this purpose. The advantage of feature selection is two-fold. It reduces the number of features thereby increasing the algorithm's speed, and it removes the unwanted or redundant features thereby increasing the efficiency of the classifier. Due to these advantages, it is recommended that researchers make use of such algorithms for better efficiency of KFE.

Once the features are extracted, then classification process is applied, which performs the actual differentiation between non-key frames and key frames. Classifiers like threshold-based methods, neural networks, machine learning methods, and others are used for this purpose. Upon classification, the post-processing layer is applied to remove any remaining outliers from the extracted key frames. This layer does the task by using simple methods like thresholding, or block difference calculations. Complex methods can be applied in this layer too, but they would reduce the algorithm's speed efficiency.

In this work, we have designed a novel region of interest evaluation algorithm that utilizes different color spaces with a visual attention model. This segmented image is given to a novel motion vector evaluation engine, wherein feature extraction and selection process is done. Finally, a neural network is applied to analyze the patterns and produce effective keyframes. The next section describes different algorithms that are studied for this purpose, followed by the proposed algorithm and its results. The paper concludes by recommending some further research in this field of KFE.

2. LITERATURE SURVEY

Over the years various researchers have developed different algorithms for KFE, some of them utilize the spatio-temporal differences between the frames, while others work on frequency components, or some other spatial domain. The work done in [1] utilizes the Pearson correlation coefficient (PCC) and the color moments (CM) between consecutive frames and evaluates the mean and standard deviation of these values. The utilization of this combo feature set enables the algorithm to perform effectively. The frames with highest values of PCC and CM mean & standard deviation are selected as key frames. Upon comparison it is found that the proposed method outperforms other methods in terms of figure of merit by atleast 20%. This method is unable to perform well for animations videos, but can perform exceptionally well for serials, personal interviews, news and movies. A video annotation method that uses KFE is proposed in [2], here the researchers have used color histogram difference (CHD) & Edge change ratio (ECR) for detection of KFE. They have used a 2-level comparison algorithm which utilizes CHD, ECR along with Fuzzy C Means in order to find out the most variant key frames from the set of input frames. They have formulated a change ratio (CR) factor in order to perform this task. The CR takes into consideration the values of CHD & ECR and then evaluates the values of mean and standard deviation from them. These values are compared frame-by-frame and the difference is termed as change ratio. Wherever there is a more than threshold change in the CR values, then those frames are termed as key-frames. The figure of merit for this technique is very high and is able to compress the video by upto 95%. The work in [3] compares different methods like SSIM (Structural Similarity Index Method) Method, Entropy Method and Euclidean Distance method, and also proposes a new method which is based on Euclidean Distance method with Differential

Evolution (DE) algorithm. The proposed method utilizes the concept of soft computing and bio-inspired computing in order to find out the key frames. The researchers have modified the existing DE algorithm by adding Euclidean distance and Entropy Difference to the fitness function. This results in obtaining key frames with higher Euclidean distance and Entropy Difference values. There is no comment on the statistical performance of the algorithm, but it should be good enough for most of the video types.

Finding out KFEs in human videos is a challenging task, because the movements are either very rapid or are very gentle. A novel method that utilizes Deep Learning to identity key frames in human videos is proposed in [4]. They have utilized the concept of deep learning in feature extraction, and then used a parallel processing engine to find out the best features suited for KFE. These features are given to a convolutional neural network (CNN) in order to find out the most varying frames. The results indicate that the proposed method is able to obtain a 95% precision and a high recall rate when compared with Discrete cosine coefficients and rough sets theory, Content relative thresholding, Multi-scale color contrast, relative motion intensity, and relative motion consistency & Color and structure features. CNN being an adaptive method is able to perform this classification with utmost accuracy. Another adaptive method is based on adaptive clustering is proposed in [5], wherein the RGB frames are converted in HSV domain, and then using adaptive clustering method the key frames are extracted. The HSV histogram is evaluated, and then a one-dimensional eigen vector is formed from this histogram. This histogram is given to a cluster validation silhouette coefficient, where density peak clustering algorithm (DPCA) is applied. The results showcase that the DPCA method outperforms K-Means, improved agglomerative hierarchical clustering algorithm (AGNES), Hierarchical Clustering and I-Frame methods.

KFE methods usually identify frames that have higher movement values, and therefore moving object detection can be a feature vector for this purpose. The work in [6] uses Moving Object Detection and Image Similarity for the purpose of KFE. It uses ViBe algorithm and fuses inter-frame difference method by dividing the original video into several segments that contain the moving object. For feature extraction Speed Up Robust Features or SURF is used. Finally, an adaptive selection threshold is used for finding out key frames from the input set of frames. They also use the concept of Peak Signal to Noise Ratio (PSNR) in order to find out frame similarity. The system first takes the video frames as input and finds out moving objects from these frames. These object frames are given to PSNR computation, and a global PSNR similarity feature is evaluated. Local similarity is found using SIFT technique, and then a weighted fusion feature set is evaluated to find the values of similarity between the frames. This similarity is compared using adaptive threshold selection to finally evaluate the key frames. Using the proposed method an accuracy of 99% can be achieved. But this is tested on a limited set of video sequences, it is recommended that the readers must perform further diligence before using this method for their own research. Another 2 level KFE method is proposed in [7], wherein researchers have used the concept of CBD and ECR for KFE. Results showcase that the proposed method has good KFE performance.

KFE in 3D videos has been a topic of study for more than a decade now. Methods like shot boundary detection (SBD), uniform sampling, position sampling, clustering (k-Means and FCM), Curve simplification, Minimum correlation, Minimum reconstruction error, Matrix factorization are studied in [8]. From the evaluation done in [8], it is found that the SBD methods outperform other methods for 3D KFE. SBD methods also use Uniform Local Binary Pattern (ULBP) in order to find out key frames. The work in [9] utilizes Uniform Local Binary Patterns between different frames of a video, and then a threshold is applied to these frames in order to find out the key frames from the given video sequence. The proposed method outperforms SIFT, Unsupervised and other methods in terms of precision, recall, F-measure, figure of merit and

accuracy. Converting video sequences into different color spaces like HSV, LAB, and others can significantly improve the KFE performance. The work done in [10] converts the frame sequences into a summary space, which is a clustered space formed after application of deep features on the input frames. This summary space is an indicative of the temporal differences between the frames. The selected weighting method is used in order to find the most relevant frames using a frame thresholding technique. The proposed approach outperforms clustering based approach, dictionary-based approach and object-based approaches when applied on open video dataset and YouTube video datasets. Another review of KFE for 3D videos is described in [11], which indicates that the performance of SBD methods is better than any of the other methods. SBD and other approaches usually use feature mapping for the purpose of KFE, these features can be histogram, color maps, or complex SURF features. The SURF features outperform other feature vectors in terms of efficiency of key frame extraction. The work in [12] utilizes the concept of SURF for feature extraction, and validates that SURF outperforms other feature extraction methods for KFE. It is recommended that SURF features be combined with deep CNNs in order to evaluate their performance as an integrated KFE algorithm. Moreover, techniques like hash map can be utilized for the purpose of KFE as described in [13]. In [13], researchers have combined hash maps with a threshold-based classifier in order to find out the key frames from given video sequence. There is no comment on the statistical performance of this algorithm, and it is recommended that researchers must evaluate the performance of this novel method before using the same. Another adaptive clustering-based method is described in [14] that indicates that the performance of HSV with adaptive clustering is superior than others for KFE.

Security in KFE has always been a secondary consideration parameter. But for highly secure applications like military and health care, it is required to have a certain level of security during KFE process. The work in [15] proposes the concept of video partitioning for securing KFE for industry standard MPEG video sequences. It again uses SBD for key frame extraction, but adds the concept of hashing and crypto-space operations for securing the KFE process. This method is able to reduce the video sequence length by more than 90%. News videos are usually long sequences that repeat information in a loop. The work done in [16] applies pixel difference between 2 consecutive frames in order to perform KFE. This pixel differencing technique is combined with text segmentation, and if the text values change, then the frames are marked as initial key frames. These initial key-frames are then given to a k-Means algorithm in order to find out the frame difference and finally obtain the key frames from the video sets. A similar work is performed in [17], that utilizes visual attention model and motion energy vectors to find out key frames. The visual attention model is devised using global similarity feature like color histogram, object shape, optical flow and other methods. Then PSNR values are evaluated in order to find out the frame similarity. Dissimilar frames are marked as initial key frames, and then are used for the second stage of processing. In the second stage, local similarity features are evaluated using Harris Corner Detection and optical flow. Based on these features, and a threshold detector the desired key frames are extracted. The results showcase that the proposed algorithm is able to find out key frames with more than 90% efficiency. Another color histogram-based method is proposed in [18], wherein the HSV model is combined with color space quantization, and a color histogram is evaluated. Then using Euclidean distance between the histograms of successive frames, the key frames are extracted. The results are evaluated on facial videos, and it is found that the rate of facial recognition improves by more than 10% when compared to other methods. A dictionary-based method is described in [19], wherein Latent Semantic Analysis (LSA) is applied along with high level detectors to evaluate key frames. Here, subtractive clustering is used in order to find out the most variant frames from the given set of input frames. The proposed work is able to improve the precision value to about 76%, which is low when compared to other methods like CNN. Thus, the LSA method can be integrated with other methods like CNN and deep CNN in order to improve the accuracy of KFE. In order to evaluate the performance of KFE methods, the work in [20-22] can be used, wherein different approaches like Pixel-Based, Block-

Based, Histogram-Based and Clustering-Based are studied. They have also studied different performance metrics like precision, recall, f-measure, figure of merit, and accuracy.

3. PROPOSED METHODOLOGY

The proposed method is based on the concept of visual attention model evaluated over various color spaces, and then combines the motion vectors evaluated from these models to form an integrated vector. This vector is a combination of different motion vector fields, and is given to a convolutional neural network for final evaluation of key frames. The following diagram indicates the overall flow of the algorithm,

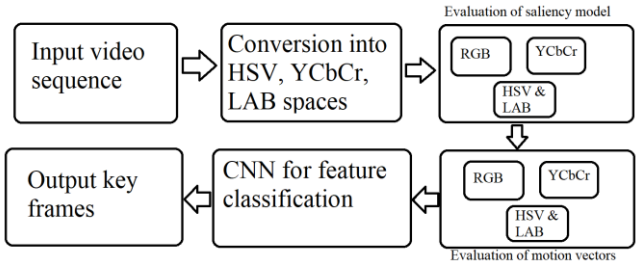

Figure 2. Flow of the proposed system

From the figure, the following flow of the system can be observed,

- The input frames are given a color space conversion block, which coverts the frames into HSV, YCbCr and LAB color spaces [22]
- Conversion into color spaces assists in finding out the variations in the color, texture and intensity of the image objects
- These different color models are then given to a saliency map detection block, where the visual saliency map [23] is evaluated
- This visual saliency map reduces those regions which are termed as backgrounds, and only foreground regions are extracted
- Due to this, the redundancy in the images is reduced, and this also helps in further improving the selection of feature vectors from the images
- Motion vectors [24] are evaluated from these saliency map images
- These vectors are formed from the pixel differences between the frames, and therefore can identify parts of the frames that are either highly moving, moderately moving or sparsely moving
- The motion vectors are evaluated for each of the saliency map frames, and the final motion vector is formed
- This motion vector is given to a CNN model, that performs classification of the frames into key frames and non-key frames
- The CNN model is trained using 200 different videos varying in terms of frame length, video content and frame size. Each of the videos have been taken from YouTube video dataset. The architecture of CNN can be seen from the following figure 3. In the architecture, convolution + ReLU layers are combined with max pooling layers for feature extraction, then a fully connected ReLU layer is combined with a softmax layer for classification. The proposed classifier is compared with other state-of-the art methods, and the results are evaluated. These results are evaluated in terms of accuracy, precision, recall, f-measure, figure of merit and compression ratio. It is found that the proposed method outperforms other non-CNN based

methods by atleast 20% and it improves the efficiency of the CNN methods by 10% on an average.

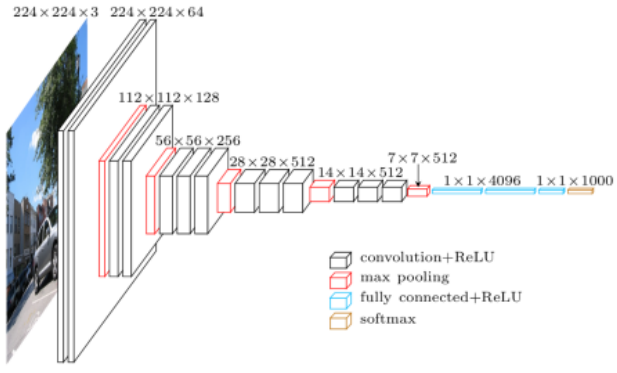

Figure 3. Used CNN architecture

The result evaluation process is described in the next section.

4. RESULTS

We compared the performance of the proposed CNN-based hybrid visual attention model using motion vector KFE technique with other methods in terms of accuracy, precision, recall, f-measure, figure of merit and compression ratio. The following results were obtained,

Table 1. Proposed results

Method	Acc. (%)	P (%)	R (%)	F	CR (%)	FOM
PCC & CM [1]	85.0	82.0	78.0	80.0	80.0	0.8
CHD & EHR [2]	91.0	90.0	93.0	91.5	76.0	0.9
ED based DE [3]	86.0	84.0	86.0	85.0	85.3	85.3
CNN [4]	94.0	91.0	83.0	86.8	89.3	88.8
HSV & DPCA [5]	91.0	89.0	81.0	84.8	87.0	86.6
ViBE & SuRF [6]	76.0	78.0	88.0	82.7	80.7	81.1
ULBP [9]	79.0	79.0	81.0	80.0	79.7	79.7
Summary Space [10]	83.0	84.0	86.0	85.0	84.3	84.5
SuRF [12]	75.0	77.0	71.0	73.9	74.3	74.2
Visual Attention with Sal Maps [16]	79.0	81.0	88.0	84.4	82.7	83.0
Optical Flow [17]	84.0	85.0	86.0	85.5	85.0	85.1
HSV with color space quant [18]	88.0	89.0	89.0	89.0	88.7	88.7
LSA [19]	76.0	77.0	78.0	77.5	77.0	77.1
Proposed [This]	98.0	97.0	94.0	95.5	96.3	96.2

From the table we can observe that the proposed system outperforms other state-of-the art methods by almost 20%. These tests were conducted on a large set of videos taken from the YouTube dataset, and each of the given algorithms were evaluated on the same dataset. The training set consists of 200 videos of different length, sizes and content, while the testing set consists of 300 videos varying in terms of length, sizes and content. It is found that the proposed algorithm performed very effectively in terms of all the parameters of evaluation.

5. CONCLUSION

From the results it is evident that the proposed model is better in terms of accuracy, precision, recall, f measure and compression ratio. These parameters are taken for more than 500 video sequences, and the final test was performed. This is due to the fact that the multi-colour saliency map when combined with motion vectors provides an effective feature vector for key frame extraction. This effective vector when combined with a high end complex classifier like CNN, is able to identify patterns for key frames, that are usually missed by linear classifiers. Due to this combination of effective segmentation, feature extraction and classification, the resultant algorithm is superior than other state of the art methods. Due to the use of CNN, the delay of training is very high (it took around 2 days for the model to get trained on a core i5 machine), thus we would recommend researchers to work on reducing the training delay for the system.

REFERENCES

[1] Reddy Mounika Bommisetty, Om Prakash2 · Ashish Khare 'Keyframe extraction using Pearson correlation coefficient and color moments' Springer-Verlag GmbH Germany, part of Springer Nature 2019.

[2] Shailendra S. Aote , Archana Potnurwar2 'An automatic video annotation framework based on two level keyframe extraction mechanism',Springer Science+Business Media, LLC, part of Springer Nature 2018.

[3] Kevin Thomas Abraham(&), Manikandan Ashwin, Darshak Sundar, Tharic Ashoor, and Gurusamy Jeyakumar 'Empirical Comparison of Different Key Frame Extraction Approaches with Differential Evolution Based Algorithms', Springer International Publishing AG 2018 S.M. Thampi et al. (eds.), Intelligent Systems Technologies and Applications, Advances in Intelligent Systems and Computing 683.

[4] Ujwalla Gawande, Kamal Hajari and Yogesh Golhar, 'Deep Learning Approach to Key Frame Detection in Human Action Videos', 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0),

[5] Hong Zhao ,Wei-Jie Wang ,Tao Wang ,Zhao-Bin Chang and Xiang-Yan Zeng, 'Key-Frame Extraction Based on HSV Histogram and Adaptive Clustering', Hindawi Mathematical Problems in Engineering Volume 2019, Article ID 5217961, 10 pages, https://doi.org/10.1155/2019/5217961

[6] Yuan Luo, Hanxing Zhou*, Qin Tan, Xuefeng Chen, and Mingjing Yun, 'Key Frame Extraction of Surveillance Video based on Moving Object Detection and Image Similarity1', *ISSN 1054-6618, Pattern Recognition and Image Analysis, 2018, Vol. 28, No. 2, pp. 225–231. © Pleiades Publishing, Ltd., 2018.*

[7] Shailendra S. Aote1 & Archana Potnurwar, 'An automatic video annotation framework based on two level

[8] keyframe extraction mechanism', Springer Science+Business Media, LLC, part of Springer Nature 2018

[9] Lino Ferreira, Luis A. da Silva Cruz and Pedro Assuncao, 'Towards key-frame extraction methods for 3D video: a review', Ferreira et al. EURASIP Journal on Image and Video Processing (2016) 2016:28 ,DOI 10.1186/s13640-016-0131-8

[10] B. Reddy Mounika, Om Prakash, Ashish Khare, 'Key Frame Extraction using Uniform Local Binary Pattern', 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), 978-1-5386-4146-0/18/$31.00 ©2018 IEEE

[11] Xuelong Li, Bin Zhao, and Xiaoqiang Lu, 'Key Frame Extraction in the Summary Space', 2168-2267 _c 2017 IEEE, IEEE TRANSACTIONS ON CYBERNETICS

[12] Lino Ferreira , Luis A. da Silva Cruz and Pedro Assuncao, 'Towards key-frame extraction methods for 3D video: a review', Ferreira et al. EURASIP Journal on Image and Video Processing (2016) 2016:28, DOI 10.1186/s13640-016-0131-8

[13] Rafa_l Grycuk, Micha_l Knop, and Sayantan Mandal, 'Video Key Frame Detection Based on SURF Algorithm', Springer International Publishing Switzerland 2015 L. Rutkowski et al. (Eds.): ICAISC 2015, Part I, LNAI 9119, pp. 566–576, 2015. DOI: 10.1007/978-3-319-19324-3_50

[14] Hariharan. K1, Arjun. S. V2, Nivedha. S3, Srithar. K4, Thivaharan. S5, 'Visual Content Based Video Indexing Using Key Frame Extraction', International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 06 Issue: 03 | Mar 2019

[15] Hong Zhao , Wei-Jie Wang , Tao Wang ,Zhao-Bin Chang and Xiang-Yan Zeng, 'Key-Frame Extraction Based on HSV Histogram and Adaptive Clustering', Hindawi Mathematical Problems in Engineering Volume 2019, Article ID 5217961, 10 pages, https://doi.org/10.1155/2019/5217961

[16] K.S.Thakrea*,A.M.Rajurkarb,R.R.Manthalkar, 'Video Partitioning and Secured Keyframe Extraction of MPEG Video', International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015,Nagpur, INDIA

[17] Harsha H Phadke, Mallika H, 'Key Frame Extraction, Localization and Segmentation of Caption Text in News Videos', 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India

[18] *Suresh C. Raikwar, '*A Framework for Key Frame Extraction from Surveillance Video', 2014 5th International Conference on Computer and Communication Technology, 978-1-4799-6758-2/14/$31.00 ©2014 IEEE

[19] Xintao Li and Tiongrong Xu, 'Face Video Key-Frame Extraction Algorithm Based on Color

[20] Histogram', 2011 International Conference on Computer Science and Information Technology (ICCSIT 2011), IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore, DOI: 10.7763/IPCSIT.2012.V51. 112

[21] Evaggelos Spyrou · Giorgos Tolias, Phivos Mylonas · Yannis Avrithis, 'Concept detection and keyframe extraction using a visual thesaurus', Multimed Tools Appl (2009) 41:337–373 DOI 10.1007/s11042-008-0237-9, Published online: 8 November 2008, © Springer Science + Business Media, LLC 2008

[22] *Hana Gharbi, Sahbi Bahroun, Mohamed Massaoudi and Ezzeddine Zagrouba, '*Key Frames Extraction Using Graph Modularity Clustering For Efficient Video Summarization', 978-1-5090-4117-6/17/$31.00 ©2017 IEEE

[23] Ciocca Gianluigi , Schettini Raimondo, 'An innovative algorithm for key frame extraction in video Summarization', J Real-Time Image Proc (2006) 1:69–88 DOI 10.1007/s11554-006-0001-1

[24] Toran Verma, Sipi Dubey, 'Impact of Color Spaces and Feature Sets in Automated Plant Diseases Classifier: A Comprehensive Review Based on Rice Plant Images', Archives of Computational Methods in Engineering, https://doi.org/10.1007/s11831-019-09364-6

[25] Alberto Lopez-Alanis1 · Rocio A. Lizarraga-Morales2 · Marco A. Contreras-Cruz1 · Victor Ayala-Ramirez1 · Raul E. Sanchez-Yanez1 · Felipe Trujillo-Romero1, 'Rule-based aggregation driven by similar images for visual saliency Detection', Applied Intelligence https://doi.org/10.1007/s10489-019-01582-6

[26] Ciocca Gianluigi Æ Schettini Raimondo, 'An innovative algorithm for key frame extraction in video Summarization', J Real-Time Image Proc (2006) 1:69–88 DOI 10.1007/s11554-006-0001-1

AUTHORS

Darshankumar Billur

Assistant in ECE Department of KLE College of Engineering & Technology Chikodi with an experience of 16 years in the teaching and research. Special interest in Image processing, Microcontrollers operating system and dedicated to contribute advancements in the field

Dr.Manu T.M.

Professor and Head of Department of Electronics & Communication Engineering at KLE Institute of Technology, Hubli with an experience of more than 25 years in the field of Teaching with with special interest in Digital Electronics, VLSI Embedded System, Image Processing and related fields.