

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800

Open access books available

122,000

International authors and editors

135M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.

For more information visit www.intechopen.com



Deep Learning Approach to Key Frame Detection in Human Action Videos

Ujwalla Gawande, Kamal Hajari and Yogesh Golhar

Abstract

A key frame is a representative frame which includes the whole facts of the video collection. It is used for indexing, classification, evaluation, and retrieval of video. The existing algorithms generate relevant key frames, but additionally, they generate a few redundant key frames. A number of them are not capable of constituting the entire shot. In this chapter, an effective algorithm primarily based on the fusion of deep features and histogram has been proposed to overcome these issues. It extracts the maximum relevant key frames by way of eliminating the vagueness of the choice of key frames. It can be applied parallel and concurrently to the video sequence, which results in the reduction of computational and time complexity. The performance of this algorithm indicates its effectiveness in terms of relevant key frame extraction from videos.

Keywords: deep learning, neural network, histogram, video processing, computer vision

1. Introduction

In video analysis and processing, relevant and necessary information retrieval is a mandatory task, because if the video is large, then it is difficult to process the complete video in less time without losing its semantic details. Key frame extraction is a primary step of a computer vision algorithm. The key frame means the part of the video that can represent a visual summary and meaningful information about the video sequence. The key frame can be useful in many applications such as video scene analysis, browsing, searching, information retrieval, and indexing. Aigrain et al. in [1] describe the benefits of key frame extraction for information extraction in a video sequence. HongJiang et al. [2] significantly justify that for any video sequence the user can perform searching, indexing, and retrieval of information efficiently and faster using key frame extraction. Liu et al. [3] and Gargi et al. [4] proposed an object motion-based approach of key frame extraction. Basically, the video has a complex structure. It is a combination of the scene, shot and frames [5] as shown in **Figure 1**. In many computer vision applications such as content-based video retrieval (CBVR), video scene analysis and video sequence summarization is mandatory to analyze the overall video structure. Video analysis major components are video scene segmentation, shot boundary detection, key frame selection, and extraction [6–8]. The main use of key frame extraction is to reduce the redundant

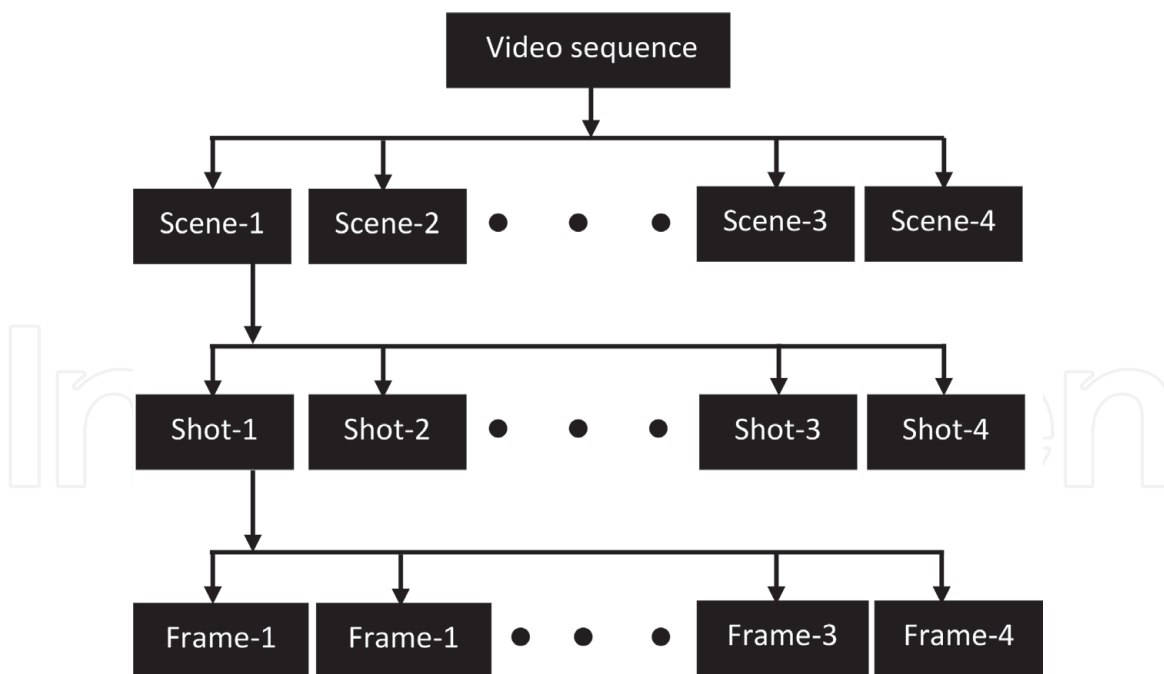


Figure 1.
Structure of video.

frames in a video to make a video scene readable and compact and prepare video sequences for faster processing.

Conventional key frame extraction methods eliminate the redundant and similar frame in a video without affecting the semantic details visual content. These techniques inputs are either a complete video or a video is divided into a set of shots by shot boundary detection methods. As shown in **Figure 1**, the shot is a consecutive, adjacent sequence of frames captured by the video camera. Thus, in this chapter we propose an efficient approach for video key frame extraction, which is faster, accurate, and computationally efficient. This chapter is organized into the following sections. Section 1 gives an introductory part of video structure and the importance of key frame extraction in a video surveillance system. Section 2 describes the existing approach for key frame size selection algorithms. Section 3 describes the existing key frame extraction methods with its issues and challenges. Section 4 describes the proposed approach for key frame extraction. Section 5 discusses experimental results and possible future directions. Finally, the chapter concluded with a discussion in Section 6.

2. Key frame size estimation methods available in the literature

The major problem we face in a key frame extraction algorithm is computing the size or number of key frames for a specific video sequence. In literature, there are several methods available for the key frame size estimation. In this section, we have discussed these methods in brief. In [3] approach the author has considered one key frame for each shot of a video. The selection of the key frame in each shot is based on the maximum entropy value of each shot. This consideration is not appropriate and accurate for the video which is having a big shot. Again, many of the useful frames of the video are discarded due to the pre-defined fixed selection of key frames. Lesser key frame extraction does not solve the problem. A set of key frames having necessary and sufficient representation of the visual content of the video is required in the output. In other proposed approaches, first, middle, or ending

location frames in the shot are considered. But the resulting frames are having a low correlation with each other in visual content. These methods are computationally less complex, but having less accuracy. In [9, 10], authors have described the three different ways of identifying the key frames in a video sequence. Each method is described in brief as follows.

2.1 Priori knowledge base as a fixed number

In this method, a pre-defined number of fixed key frames are considered as a fixed value before the key frame extraction process. Consider “k” as the number of key frames, and then the key frame set K_r is defined by Eq. (1):

$$K_r = \{F_{i1}, F_{i2}, F_{i3}, \dots, F_{ik}\} \quad (1)$$

The sequence of video frames is the change as per the type of video. The specific summarization of key frames is defined by Eq. (2):

$$\{K_{f1}, K_{f2}, K_{f3}, \dots, K_{fk}\} = \sum_n^1 \min_{ri} \{Dist(K_r, V, \delta)\} \quad (2)$$

where,

$1 \leq r_i \leq n$ and

n is represented as a number of frames in a video, δ represents the key frame summarization factor, and $Dist$ represents the distance measure, i.e., used for computing dissimilarity between frames. The δ in this method is useful for maintaining a lesser number of key frames by covering complete visual content details in the video.

2.2 Posteriori knowledge base as unknown

In this method, the number of key frames is not fixed. The number of key frames is unknown until the key frame extraction process gets completed. The key frame size is depending upon the type of content of the video frame. If the video scene consists of dynamic action movements, then the number of key frames is more otherwise less for static video scenes. Key frame generation can be represented by Eq. (3):

$$\{K_{f1}, K_{f2}, K_{f3}, \dots, K_{fk}\} = \sum_n^1 \min_{ri} \{K | Dist(K_r, V, \gamma)\} \quad (3)$$

where γ parameter is used for tolerance to dissimilarity level. Another parameter is similar to the previous method.

2.3 Determined-fixed number

In this method, the number of key frames is predetermined before the whole process key frame extraction process. In [11, 12] approaches key frames are extracted using the clustering technique. The key frame extraction algorithms stop when extracted key frame size matched with the pre-defined key frame value.

3. Key frame extraction methods with its issues and challenges

In literature, there are several methods to extract key frames. Hannane et al. [13] and Hu et al. [14] categorize the key frame extraction into different categories as a sequential comparison of frames, global comparison of frames, the minimum correlation between frames, minimum reconstruction error in frames, temporal variance between frames, maximum coverage of video frames, reference key frame, curve simplification, key frame extraction using clustering, object- and event-based key frame extraction, and panoramic key frames. Each of these methods is described in brief as follows.

3.1 Sequential comparison of frames

In this method, each frame of a video sequence is compared with the previously extracted key frame. If the difference between the extracted key frame and the current key frame is high, then this frame is considered as the new key frame. In [13] key frames are extracted based on the color histogram comparison between the current and previous frames of a video sequence. The main advantage of this method is that it is simple and computationally less complex. But the disadvantage is that the extracted key frame consists of redundant key frames.

3.2 Universal frame comparison

In this method, the global difference between frames in a shot is computed using a predetermined objective function, which is application-specific. Zhuang et al. in [9] describes the different objective functions for comparison of frames in the shot. Each of these functions is discussed in brief as follows.

3.3 Minimum associations

In this method, relevant key frames are generated from a shot by reducing the summation of the association between frames. The extracted key frames are tightly coupled with each other. Liu in [3] uses a graph-based approach to extract distinct key frames with their association. Weight directed graph is used to represent the shot, and the shortest path is computed using the A* algorithm. The frames, which are having minimum association and less correlation, are represented as key frames in the shot.

3.4 Minimum reformation error

In this method, the key frames are extracted by reducing the variation between the prevision frame and set of frames in a shot. The prevision frame is generated by the numeric analysis method interpolation. Chao et al. in [15] presented an approach to select a pre-defined set of key frames and reduce the frame reformation error. In [16] a combined approach of the prevision frame-based approach and a pre-defined set of key frame selection approach is proposed. This method uses the motion-based features.

3.5 Similar temporal variance

In these methods, frames having similar variance are selected as the key frames of the specific shot [17]. The sum of temporal variance between all frames is

selected as an objective function. The temporal variance is computed by the summation of change in the frame content in a shot.

3.6 Maximum key frame representation coverage

In this method, the representation coverage of a key frame means a number of frames in a shot that a key frame can cover [18]. This method can be useful in the size of the key frame selection. The advantage of this method over a universal comparison method is that the extracted key frames are maintainable and consist of global context information of a shot. The only disadvantage of this method is that it is computationally complex.

3.7 Predetermined reference frame

In this method, a key frame is generated by comparing the predetermined reference frame and each frame in a shot [19]. The main advantage of this method is that it is not computationally complex and easy to implement. Its drawback is that it does not represent the global context in a shot efficiently.

3.8 Trajectory curve simplification

In this method, the trajectory curve is generated from the frames. The curve consists of a sequential combination of points in the feature. Calic and Izquierdo in [20] presents a dynamic method for change detection in the scene and the key frame generation. The frame difference metric is computed using the small size block features in a scene. After that contour detection method is used for trajectory curve plotting using the metric.

3.9 Cluster-based key frame extraction

In this method, key frame clusters are created using the data points and features of video sequences. The set of key frames is created with frames that have the closest distance from the center of the cluster. In [21, 22] fuzzy K-means- and fuzzy C-means-based methods for the key frame selection are presented. The clusters are generated based on the different features like motion sequences and the distance matrix score. In [23] an approach that combined K-means and mean squared error for the key frame selection is presented. Pan et al. in [24] proposed an enhanced fuzzy C-means clustering algorithm for the key frame selection. Clusters are generated using the color feature. The key frames having the highest entropy are considered as a key frame from each cluster. The advantage of cluster-based approaches is that it covers the global characteristics of the scene. The disadvantage of these methods is that it requires a high computational cost for cluster generation and feature extraction from the scene.

3.10 Event-driven key frame extraction

In this method, the extracted key frame consists of event and object details. The advantage of this method is that each key frame describes the object motion pattern, object, and event details [25]. The drawback of this method is that the pre-defined rules need to be defined as per the application, identifying objects and events in a key frame. Hence, the accuracy of this algorithm depends upon the pre-assumption parameters set before the key frame extraction algorithm is executed.

3.11 Full details key frame extraction (panoramic frame)

In this method, the key frame consists of the complete detail of a scene in a shot. Papageorgiou and Poggio in [25] presented a key frame extraction approach using

| Method name | Characteristics | Advantage | Shortcomings of the method | Ref. | Year |
|--|---|--|--|------|------|
| Clustering method (Zhuang et al.) | Analysis of short boundary video | Faster processing | <ul style="list-style-type: none"> • Less key frame selection for single-shot activity • More key frame selection for multiple | [9] | 1998 |
| Entropy method (Mentzelopoulos et al.) | Best method for unpredictable dataset | Local feature selection | <ul style="list-style-type: none"> • External effects such as lighting condition affect the performance | [10] | 2012 |
| Histogram method (Rasheed et al.) | Similarity measure between key frames | High-level segmentations | <ul style="list-style-type: none"> • Cannot consider the local similarities | [11] | 2015 |
| Motion analysis method (Wolf et al.) | Optical flow-based analysis | Faster mid-range key frame selection | <ul style="list-style-type: none"> • Highly depends on the static frame references | [12] | 2016 |
| Triangle-based method (Liu et al.) | Determination of the motion characteristics | Reduces the motion effects on the video | <ul style="list-style-type: none"> • Cannot detect the color-based information change | [3] | 2016 |
| 3D augmentation method (Chao et al.) | Processing short and fast motion video data | Combines the video data into multidimensional model | <ul style="list-style-type: none"> • Highly time complex | [15] | 2018 |
| Optimal key frame selection method (Sze et al.) | Best method for continuously growing video sequence by adopting the temporary key frame | Faster processing due to probabilistic analysis | <ul style="list-style-type: none"> • Highly time complex | [16] | 2017 |
| Context-based method (Chang et al.) | Best method for repetitive information contents | Generates a multilevel abstract of the information | <ul style="list-style-type: none"> • Information loss due to less key frame selection | [17] | 2017 |
| Motion-based extraction method (Luo et al.) | Adopts the advantages from digital capture devices | Reduces the spatiotemporal effects | <ul style="list-style-type: none"> • High-quality video information expected | [18] | 2015 |
| Robust principal component analysis method (Dang et al.) | Adopts the decomposition method for sparse component analysis | Analyzes the frames for consumer videos with fewer contents or rapid content shift | <ul style="list-style-type: none"> • Assumptions are not always reflecting better results | [19] | 2010 |

Table 1.
Recently used pedestrian databases by the researchers.

the homography matrix. The main advantage of this method is that it covers the global context of the shot. The drawback of this method is that it is having high computational complexity. The comparative analysis of recently utilized key frame extraction algorithms is shown in **Table 1**. The comparison is performed in terms of characteristics, advantages, and shortcomings of the method.

4. Proposed methodology for key frame extraction

The proposed approach is based on the combination of the histogram and deep learning to extract the relevant key frame from the video sequence. **Figure 2** shows the main steps of the proposed framework. The steps of key frame extraction include (1) video reading from the database, (2) frame extraction from video, (3) preprocessing, (4) histogram generation, (5) comparison of the histogram, (6) distinct key frame generation, and (7) key frame extraction using convolution neural network (CNN). Each of these steps is described in subsequent subsections.

4.1 Video reading from database

We have tested this algorithm on the various publicly available datasets and on our own behavioral dataset. The first step is to read a video from the database. The raw video sequence selected from database is represented by Eq. (4):

$$V_i = \{V_1, V_2, V_3, \dots, V_k\} \quad (4)$$

where $1 \geq V \leq k$.

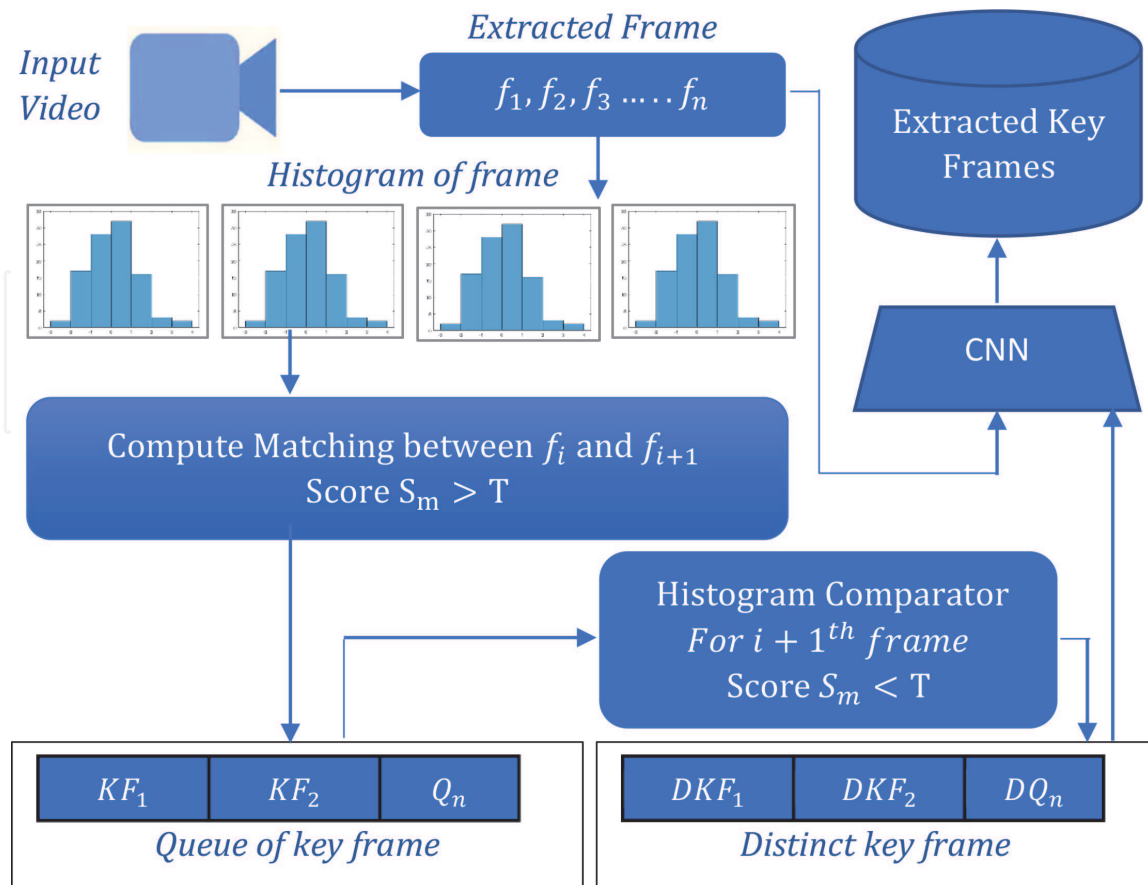


Figure 2.
 Proposed framework for key frame extraction.

4.2 Frame extraction from video

The number of frames is extracted from the video selected in step 1. The extracted frames are stored in a local directory for further processing. It is represented by Eq. (5):

$$F_i = \{F_1, F_2, F_3, \dots, F_n\} \quad (5)$$

where $1 \geq F \leq n$.

4.3 Preprocessing of frames

In the preprocessing step, the key frame queue initialized with $Q_k = 0$. The key frame queue Q_k initialized with zero because in the initial step key frame is zero. Next, the extracted frames of step 2 are converted from RGB model space to the HSV model space. This conversion is necessary to get a more specific color, gray shade, and brightness information. In HSV model space, hue is the color portion of the model, expressed as a number from 0 to 360. Saturation describes the amount of gray in a particular color, ranging from 0 to 100%. The value component represents the intensity of the color, ranging from 0 to 100%, where 0 is completely black and 100 is the brightest and reveals the most color.

4.4 Histogram generation

In this step, the normalized histogram is generated from the hue-saturation and value component in order to compare the adjacent frame. The normalized histogram is generated for contrast enhancement and compact representation of intensity and color information of the frame. Normalized histogram H_n is computed by Eq. (6)

$$H_n = \frac{\text{number of pixels with intensity } n}{\text{total number of pixels}} \quad (6)$$

where, n indicates possible intensity value.

4.5 Histogram comparison

In this step, the normalized histogram H_n is generated for each frame, and adjacent frame histogram is compared using the Bhattacharyya distance measure. It is defined by Eq. (7):

$$d(Hn_1, Hn_2) = \sqrt{1 - \frac{1}{\sqrt{Hn_1 Hn_2 N^2}} \sum_i \sqrt{Hn_1(F_p) \cdot Hn_2(F_c)}} \quad (7)$$

where:

Hn_1 indicates histogram of the previous frame F_p .

Hn_1 indicates histogram of the current frame F_c .

N indicates the number of histogram bins

The Bhattacharyya distance $d(Hn_1, Hn_2)$ is the result of a comparison of the matched score (S_m). The S_m value ranges from 0 to 1. The value 0 indicates an exact match of the content of the video frame, 0.5 is half match and 1 represents mismatch. Next, different conditions are checked to match to extract dissimilar

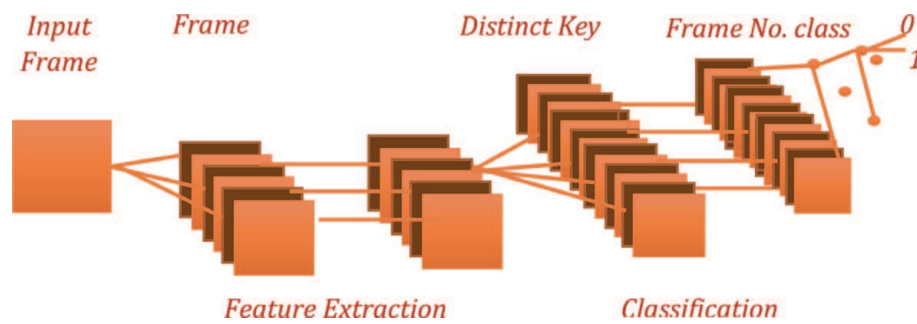


Figure 3.
 A CNN for proposed key frame extraction algorithm.

frames and similar frames. The different conditions of the score (S_m) are compared with a threshold (T) as:

1. *if* ($S_m > T$) *then* current frame is dissimilar than the previous
2. *if* ($Q_k \neq 0$) *then*
3. Add a frame in the queue of key frame $Q_k \leftarrow f_i$

4.6 Distinct key frame generation

In this step, the distinct key frame is selected, and redundant key frames are removed from the Frame queue as follows:

1. for each frame FQ_i in Q_k
2. $S_m \leftarrow (FQ_i, F_i)$
3. *if* ($S_m < T$) *then* current frame is dissimilar than the previous.
4. Add a frame in the queue of key frame $DQ_k \leftarrow f_i$.

4.7 Key frame extraction using a convolution neural network

CNN is composed of two basic parts of feature extraction and classification. Feature extraction includes several convolution layers followed by max-pooling and an activation function. The classifier usually consists of fully connected layers as shown in **Figure 3**.

Extracted distinct key frames are used as testing queries in classification phase, and input frames features are extracted using the CNN feature extraction module, and learn features are matched with distinct key frame features to obtain the best match frame which is considered as key in the output as a frame index number. The key frame extraction and CNN approach perform in parallel to obtain efficiency.

5. Experiment results and discussion

In this section, we have evaluated the efficiency of the proposed method on a publicly available database and our own human action database.

| Data source | Purpose | # Image or video clips | Annotation | Environment | Ref. | Year |
|-------------------------------|--|--|--|---|--------------|---------------|
| MIT | City street pedestrian segmentation, detection, and tracking | 709 pedestrian images, 509 training and 200 test images | No annotated pedestrian | Daylight scenario | [26] [27] | 2000, 2005 |
| Caltech Pedestrian dataset | Detection and tracking of pedestrian walking on the street | 250,000 frames (in 137 approximately minute-long segments) | 350,000 bounding boxes and 2300 unique pedestrians were annotated | Urban environment | [27] | 2012 |
| GM-ATCI | Rear-view pedestrian segmentation, detection, and tracking | 250 video sequences | 200 K annotated pedestrian bounding boxes | Dataset was collected in both day and night scenarios, with different weather and lighting conditions | [28] | 2015 |
| Daimler | Detection and tracking of pedestrian | 15,560 pedestrian samples, 6744 negative samples | 2D bounding box overlap criterion and float disparity map and a ground truth shape image | Urban environment | [29] | 2016 |
| NICTA 2016 | Segmentation, pose estimation, learning of pedestrian | 25,551 unique pedestrians, 50,000 images | 2D ground truth image | Urban environment | [30] | 2016 |
| MS COCO 2018 | Object detection, segmentation, key point detection, DensePose detection | 300,000, 2 million instances, 80 object categories | 5 captions per image | Urban environment | [31] | 2018 |
| Mapillary Vistas dataset 2017 | Semantic understanding street scenes | 25,000 images, 152 object categories | Pixel-accurate and instance-specific human annotations for understanding street scenes | Urban environment | [32] | 2017 |
| MS COCO 2017 | Recognition, segmentation, captioning | 328,124 images, 1.5 million object instances | Segmented people and objects | Urban environment | [33] | 2017 |
| MS COCO 2015 | Recognition, segmentation, captioning | 328,124 images, 80 object categories | Segmented people and objects | Urban environment | [34] | 2015 |
| ETH | Segmentation, detection, tracking | Videos | The dataset consists of other traffic agents such as different cars and pedestrians | Urban environment | [35] | 2010 |
| TUD-Brussels | Detection, tracking | 1092 image pairs | 1776 annotated pedestrian | Urban environment | [33] | 2009 |

| Data source | Purpose | # Image or video clips | Annotation | Environment | Ref. | Year |
|--------------------------------------|---|----------------------------------|---|-------------------|------|------|
| INRIA | Detection, segmentation | 498 images | Annotations are marked manually | Urban environment | [34] | 2005 |
| CVC-ADAS | Detection, tracking | 60,000 frames | 7900 annotated pedestrians | Urban environment | [35] | 2009 |
| PASCAL VOC 2012 | Detection, classification, segmentation | 11,530 images, 20 object classes | 27,450 ROI annotated 6929 segmentations | Urban environment | [36] | 2012 |
| Pedestrian behavior dataset (own DB) | Pedestrian behavior recorded in the college environment | 50 human behavior datasets | No annotated pedestrian | Daylight scenario | — | — |

Table 2.
Pedestrian databases used for the experiment for key frame extraction.

| Type of features | Recall | Precision | CPU time (ms) |
|--|--------|-----------|---------------|
| Proposed key frame extraction algorithm | 0.95 | 0.92 | 0.50 |
| Discrete cosine coefficients and rough sets theory based [1] | 0.88 | 0.82 | 0.90 |
| Content relative thresholding technique based [2] | 0.80 | 0.81 | 0.80 |
| Multi-scale color contrast, relative motion intensity, and relative motion consistency based [3] | 0.83 | 0.80 | 0.90 |
| Color and structure feature based [4] | 0.80 | 0.86 | 0.98 |

Table 3.
Comparative analysis of mean, recall, and precision and CPU time achieved by different techniques.

The results demonstrate significant improvement over the conventional methods and with low time complexity. Next, in subsequent sections, the various experiments conducted are discussed as follows.

5.1 Dataset analysis

The performance of a key frame extraction technique was evaluated and compared with the state-of-the-art methods using benchmark databases. We have taken sample videos of benchmark database and human action database as shown in **Table 2**.

5.2 Computational complexity of the proposed system

The proposed methodology is clearly superior to the rest of the techniques for key frame extraction as shown in **Table 3**. The comparative analysis of recall and precision metric for each video sequence is shown in **Figure 4**. It is observed that the proposed approach of key frame extraction achieves the highest values for recall and precision for all the video sequences. A maximum value of one of the metrics is generally not sufficient. The precision metric is used to measure the ability of a technique to retrieve the most precise results. A high value of precision means better relevance between the key frames. However, a high value of precision can be

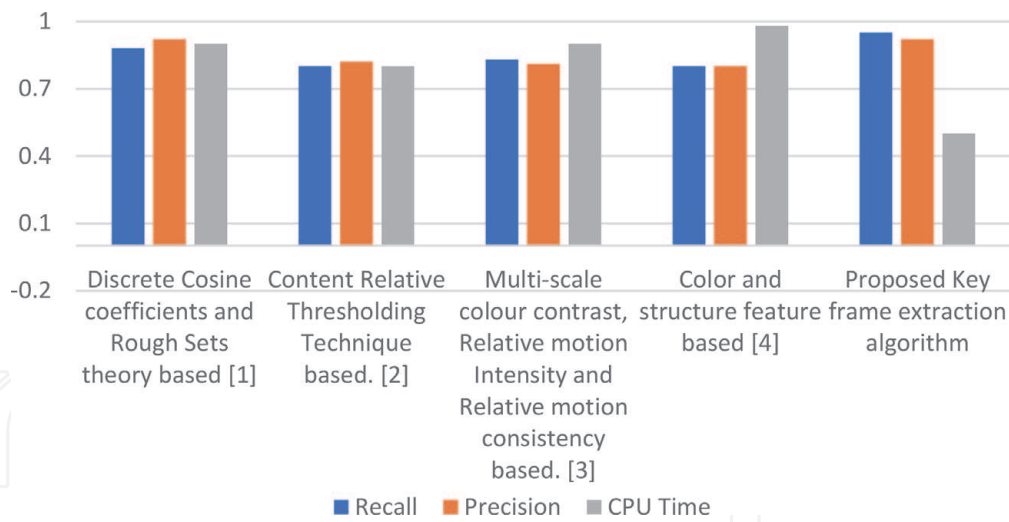


Figure 4. Recall (R), precision (P), and computational time achieved by different techniques on video dataset of **Table 2**.



Figure 5. Input video frame.

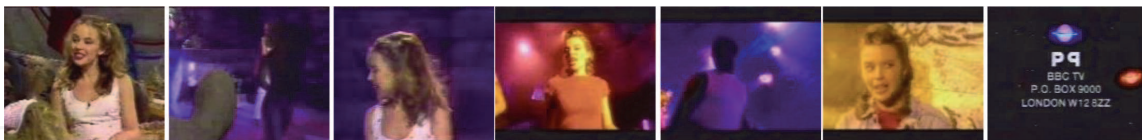


Figure 6. Extracted key frame from video.

achieved by selecting very few key frames in a video sequence. The speed and accuracy of both parameters are important in the key frame extraction algorithm. If the algorithm is slow, then the throughput of the system gets affected. It is also necessary that extracted key frames are relevant and accurate. Further, it will affect the other process, such as object detection, classification, object description, etc., respectively (**Figures 5 and 6**).

5.3 Qualitative result of frame extraction

Qualitative results from the proposed deep learning approach for the key frames extraction algorithm are shown in **Figure 7**. The figure illustrates the relevant and non-redundant key frames are extracted from the video sequence. The dataset consists of 7 suspicious student behavior. The pedestrian behaviors are recorded at prominent places of the college in different academic activities.



Figure 7.
Qualitative results of the proposed key frame extraction method on a sample video of student pedestrian dataset. (a) Frames extracted from sample video of dataset (First three-column). (b) Key frames extracted from sample video (Forth column).

6. Conclusions

This chapter describes and evaluates the methodologies, strategies, and stages involved in video key frame extraction. It also analyzes the issue and challenges of each of the key frame extraction methods. Based on the literature survey, most of the available techniques proposed by the earlier researchers can perform key frame extraction. However, most of them failed to encounter the trade-off problem between accuracy and speed. The proposed framework and approach give significant improvements for key frame extraction irrespective of the video length rather on the content type. This is made possible due to the histogram-based comparison of video scene content and convolution neural network-based deep features approach. With significantly satisfactory results, this work can generate a key frame dynamically from any video sequence. We have performed experiments on the publicly available database and obtained encouraging results.

Author details


Ujwalla Gawande^{1*}, Kamal Hajari¹ and Yogesh Golhar²

¹ Department of Information Technology, Yeshwantrao Chavan College of Engineering, Wanadongri, Maharashtra, India

² Department of Computer Science and Engineering, G.H. Rasoni Institute of Engineering and Technology, Nagpur, Maharashtra, India

*Address all correspondence to: ujwallgawande@yahoo.co.in

IntechOpen

© 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>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Aigrain P, Zhang H, Petkovic D. Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications*. 1996;**3**(3):179-202
- [2] HongJiang Z, Wang JYA, Altunbasak Y. Content-based video retrieval and compression: A unified solution. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*; October 26–29, 1997; Santa Barbara, CA. pp. 13-16
- [3] Liu T, Zhang H-J, Qi F. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*. 2003;**13**(10):1006-1013
- [4] Gargi U, Kasturi R, Strayer SH. Performance characterization of video-shot-change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*. 2000; **10**(1):1-13
- [5] Liu G, Zhao J. Key frame extraction from MPEG video stream. In: *Second symposium International Computer Science and Computational Technology (ISCSCT'09)*; December 26–28, 2009; Huangshan, P.R. China. pp. 007-011
- [6] Gawande U, Golhar Y. Biometric security system: A rigorous review of unimodal and multimodal biometrics techniques. *International Journal of Biometrics (IJBM)*. 2018;**10**(2):142-175
- [7] Gawande U, Golhar Y, Hajari K. Biometric-based security system: Issues and challenges. In: *Intelligent Techniques in Signal Processing for Multimedia Security*. *Studies in Computational Intelligence*, Vol. 660; October, 2017; Cham: Springer. pp. 151-176
- [8] Gawande U, Zaveri M, Kapur A. A novel algorithm for feature level fusion using SVM classifier for multibiometrics-based person identification. *Applied Computational Intelligence and Soft Computing*. 2013; **2013**(9):1-11
- [9] Zhuang Y, Rui Y, Huang TS, Mehrotra S. Adaptive key frame extraction using unsupervised clustering. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*; October 7, 1998; Chicago, IL, USA, pp. 866-870
- [10] Mentzelopoulos M, Psarrou A. Key frame extraction algorithm using entropy difference. In: *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR*; 15–16 October, 2004; New York, NY, USA. pp. 39-45
- [11] Rasheed Z, Shah M. Detection and representation of scenes videos. *IEEE Transactions on Multimedia*. 2005;**7**(6): 1097-1105
- [12] Wolf W. Key frame selection by motion analysis. In: *Proceedings of IEEE International Conference on Acoustics, Speech Signal Processing*; May 9, 1996; Atlanta, GA, USA, pp. 1228-1231
- [13] 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. *International Journal of Multimedia Information Retrieval*. 2016;**5**(2):89-104
- [14] Weiming H, Xie N, Li L, Zeng X, Maybank S. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 2011;**41**(6):797-819
- [15] Chao G, Tsai Y, Jeng S. Augmented 3-D keyframe extraction for surveillance videos. *IEEE Transactions*

on Circuits and Systems for Video Technology (TCSVT). 2010;**20**(11): 1395-1408

[16] Sze KW, Lam K-M, Qiu G. A new key frame representation for video segment retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*. 2005;**15**(9): 1148-1155

[17] Chang HS, Sull S, Lee SU. Efficient video indexing scheme for content-based retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*. 1999;**9**(8): 1269-1279

[18] Luo J, Papin C, Costello K. Towards extracting semantically meaningful key frames from personal video clips: From humans to computers. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*. 2009; **19**(2):289-301

[19] Dang C, Radha H. RPCA-KFE: Key frame extraction for video using robust principal component analysis. *IEEE Transactions on Image Processing (TIP)*. 2015;**24**(11):3742-3753

[20] Calic J, Izquierdo E. Efficient key-frame extraction and video analysis. In: *Proceedings of International Conference on Information Technology: Coding and Computing*; April, 2002; Las Vegas, NV, USA, pp. 28-33

[21] Nasreen A, Roy K, Roy K, Shobha G. Key frame extraction and foreground modelling using K-means clustering. In: *International Conference on Computational Intelligence, Communication Systems and Networks (CICSYN)*; Latvia; 2015. pp. 141-145

[22] Yu XD, Wang L, Tian Q, Xue P. Multilevel video representation with application to keyframe extraction. In: *Proceedings Multimedia Modelling Conference*, Australia; 2004. pp. 117-123

[23] Zhang Q, Yu S-P, Zhou D-S, Wei X-P. An efficient method of key-frame extraction based on a cluster algorithm. *Journal of Human Kinetics*. 2013;**39**(1): 5-13

[24] Pan R, Tian Y, Wang Z. Key-frame extraction based on clustering. In: *Proceedings of IEEE Progress in Informatics and Computing*; December, 2010; China. pp. 867-871

[25] Papageorgiou C, Poggio T. A trainable system for object detection. *International Journal of Computer Vision*. 2000;**38**(1):15-33

[26] Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012;**34**(4): 743-761

[27] MIT Pedestrian Dataset. Center for Biological and Computational Learning at MIT and MIT. 2005. Available from: <http://cbcl.mit.edu/software-datasets/PedestrianData.html> [Accessed: 22 September 2018]

[28] Levi D, Silberstei S. Tracking and motion cues for rear-view pedestrian detection. In: *18th IEEE Intelligent Transportation Systems Conference (ITSC)*; September 15–16, 2015; Spain. pp. 664-671

[29] Li X, Flohr F, Yang Y, Xiong H, Braun M, Pan S, et al. A new benchmark for vision-based cyclist detection. In: *IEEE Intelligent Vehicles Symposium (IV)*, Gothenburg, Sweden, June 19–22, 2016. pp.1028-1033

[30] Campbell D, Petersson L. GOGMA: Globally-Optimal Gaussian Mixture Alignment. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, IEEE; June, 2016

[31] Pellegrini S, Ess A, Van Gool L. Wrong turn – no dead end: A stochastic

pedestrian motion model. In: International Workshop on Socially Intelligent Surveillance and Monitoring (SISM'10), in Conjunction with International Conference on Computer Vision and Pattern Recognition (CVPR); June 13–18, 2010; San Francisco, CA, USA

[32] Wojek C, Walk S, Schiele B. Multi-cue onboard pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 20–25, 2009; Miami, Florida, USA

[33] Dalal N. Finding people in images and videos. GRAVIR - IMAG - Graphisme, Vision et Robotique, Inria Grenoble - Rhône-Alpes, CNRS - Centre National de la Recherche Scientifique [PhD thesis]; 2006

[34] Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*. 2015;**111**(1):98-136

[35] CVC-ADAS Pedestrian dataset. 2012. Available from: <http://adas.cvc.uab.es/site/> [Accessed: 22 September 2018]

[36] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: *European Conference on Computer Vision*; February, 2015 Springer; 2014. pp. 740-755