Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, July, 2021: 8746- 8758

Research Article

Multi-view Keyframe Extraction Techniques: A Comparative Analysis for Closed Room Scenario

Vishal Parikh, Priyanka Sharma, Chirag Patel, Henil Patel
Institute of Technology, Nirma University, Ahmedabad, India

Abstract

The enormous library of videos is developing by the day in modern era. Analyzing such vast amounts of data is often a time-consuming procedure. Access to user-friendly information is an important part of making effective use of video content. This contributes to the development of the field of study known as video summarization. This paper introduces various methods of extracting keyframes from a large video and by embedding the keyframes in a temporal graph, a summarized video would be generated.

Keywords: Multi-view video; video summarization; keyframe extraction; object detection

1. Introduction

Rapid advancements in several parts of computer infrastructure, like improved processing power, larger and less expensive storage capacity, and quicker networks, have fueled the digital video revolution in recent years [1]. Abstraction approaches are primarily intended to make storing and browsing of a video database easier. They supplement the automatic method to video retrieval (i.e. searching), especially when content-based indexing and retrieval of video sequences has had little success [2][3][4]. The main drawback of video abstraction is that it can solely be fruitful if the number of choosen video sequences is minimal [5][4]. Looking into a large number of videos to identify a particular sequence can be very time consuming and also exhausting for the user. There is no approach that we are aware of that is expressly developed to abstract films with overlapping views into multi-keyframe. Multi-keyframe, on the other hand, is a type of overlapping view video abstraction that allows users to absorb video content more quickly and intuitively when structured into a spatial shape such as a storyboard [6][7][8].

The ability to integrate audio and motion features, which may increase both the information and fluency of the abstract, is one major advantage of a video skim over a keyframe set. Furthermore, watching a skim rather than a slideshow of keyframes is often more enjoyable and interesting. However, because keyframes are not constrained by time or synchronisation difficulties, there are more options for grouping them for browsing and navigation than to the strict sequential presentation of video skims once they've been removed. For many video analysis and retrieval applications, keyframes can also help reduce computational complexity [20][4][21]. Although video skims and keyframes are frequently generated in distinct ways, it is easy to convert between the two types of video abstract. Video skims can be constructed from keyframes by combining fixed-size segments, sub-shots, or full shots that surround them. The keyframe set, on the other hand, can be constructed from the video skim by uniform sampling or picking one frame from each skim snippet.

As seen in Fig. 1 of a camera network, there are multiple non-overlapping as well as overlapping Fields Of View (FOV). Due to the overlapping field of view, the need for processing out correlating information from multiple views to produce a multi-view summary [22].

The section 2 discusses related work discussed in the literature, section 3 discusses the proposed methodology, and section 4 discuss about the result analysis of our proposed methods with existing state of the art methods.

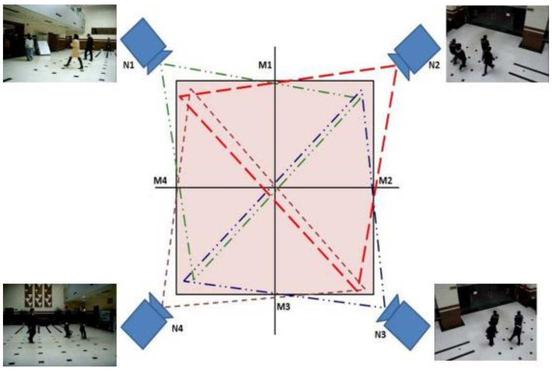


Figure 1: Camera network for multi-view scenario

2. Related Works

On summarising videos in the form of a keyframe sequence or a video skim, there is a large body of knowledge in multimedia and computer vision.

2.1. Single View Video Summarization

Much work has been made in developing unsupervised methods for summarising single-view videos as well as supervised algorithms [23][24]. Clustering, attention modelling, saliency based linear regression model, super frame segmentation, kernel temporal segmentation, crowd-sourcing, energy minimization, storyline graphs, submodular maximisation, determinant point process, archetypal analysis, long short term memory, and maximal biclique finding are some of the strategies investigated [5][25][26][23][27]. Since the reconstruction error and sparsity terms neatly fit into the summarization problem, there has been a growing interest in employing sparse coding (SC) to solve the challenge of video summarization. A novel multi-view summarising approach that jointly summarises a series of videos to obtain a single summary for representing the collection as a whole, in contrast to previous efforts that can only summarise a single video is listed in [6][28][26][27][29]. Parikh at el. in [42] compares various key-frame extraction techniques for single view video summarization for a closed room scenario.

2.2. Multi-view Video Summarization

Due to the unavoidable subject diversity and content overlaps within multi-view videos, creating a summary from them is a more difficult challenge than creating one from a single video. There have been some explicitly built algorithms that use random walk over spatiotemporal graphs and rough sets to summarise multi-view videos to solve the issues found in multi-view environments [4][6][15]. To overcome the problem of multi-view video summarization, a recent study used bipartite matching limited optimum path forest clustering [40]. An online method can also be found in [20]. However, in uncontrolled conditions, this system relies on inter-camera frame correspondence, which might be a tough challenge to solve. The work in similarly tackles a similar summarising challenge in non-overlapping camera networks [30]. A current trend in multiple online video summarization is to learn from many information sources such as video tags, topic-related web videos, and non-visual data [14]. Parikh at el. in [42] suggests camera placement strategy for generating overlapping view for multi-view video summarization.

3. PROPOSED METHODOLOGY

In this section, we provide our methodology for obtaining a multi-view summarized video using keyframe extraction techniques.

The objective of keyframe extraction is achieved by the probability of a person in each frame and the number of persons present in that frame. For the person detection in each frame we have used our own trained CNN (Convolutional Neural Network) model. For the person detection, first of all each and every frame is extracted using video processing tools and then passed through the CNN model. Frame set F contains all the frames $f1, f2, f3, ..., fn$ of a video.

$$F = f1, f2, f3, ..., fn$$

A bounding box will be created across the part of each frame where a person is detected and along with that box a probability value will be attached with it. Fig. 2 shows the model summary.

As we pass each and every frame of frame set F, we have also attached a score value corresponding to that frame. Score value represents the summation of all probability values generated in each frame of set F. We created another set S for storing the score value of all frames.

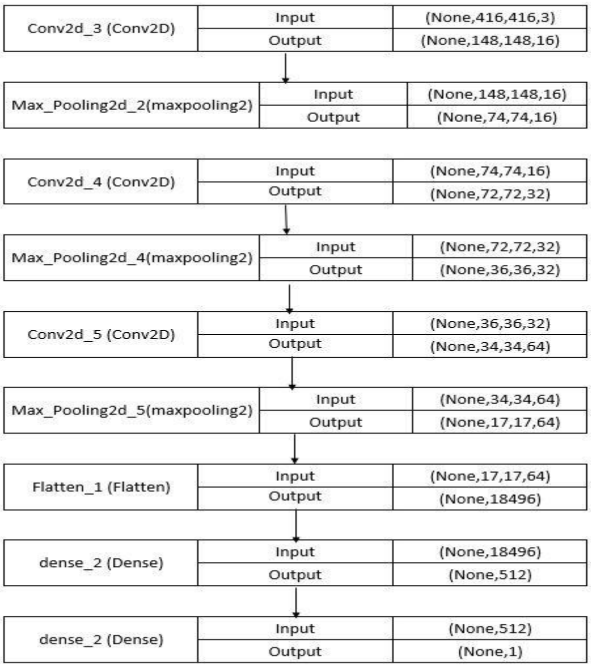
$$S = s1, s2, s3 ,..... ,sn$$

where $s1,s2,s3,......,sn$ shows the score value of frames $f1, f2, f3,......, fn$

For the keyframe extraction task we have mentioned various strategies ahead in the paper. After applying a keyframe strategy, set KF will be generated containing all the keyframes generated from set F.

$$KF = KF1, KF2, KF3,....., KFm$$

where m is the total number of key frames and n is the total number of frames $m < n$.



Layer		
Conv2d_3 (Conv2D)	Input	(None,416,416,3)
	Output	(None,148,148,16)
Max_Pooling2d_2(maxpooling2)	Input	(None,148,148,16)
	Output	(None,74,74,16)
Conv2d_4 (Conv2D)	Input	(None,74,74,16)
	Output	(None,72,72,32)
Max_Pooling2d_4(maxpooling2)	Input	(None,72,72,32)
	Output	(None,36,36,32)
Conv2d_5 (Conv2D)	Input	(None,36,36,32)
	Output	(None,34,34,64)
Max_Pooling2d_5(maxpooling2)	Input	(None,34,34,64)
	Output	(None,17,17,64)
Flatten_1 (Flatten)	Input	(None,17,17,64)
	Output	(None,18496)
dense_2 (Dense)	Input	(None,18496)
	Output	(None,512)
dense_2 (Dense)	Input	(None,512)
	Output	(None,1)

Figure 2: Model Summary

For multiple view, there would be multiple frame sets, if we have 3 views frame sets F1, F2 and F3 are generated and from these frame sets, keyframe sets KF1,KF2 and KF3 are generated. After generating 3 keyframe sets, a final set KF will be generated, after applying sorting techniques on the 3 keyframe sets, and all the frames of KF set will be joined together by video processing tools and a multi-view summarized video will be generated.

As discussed earlier, keyframes are to be extracted from the frame sets of multiple views and after applying video processing tools, a summarized video will be generated from those frames.

3.1. Maximum Frame Coverage

In this method of keyframe extraction, the main focus is to maximize the frame coverage i.e. if in any frame a person or more than one person is captured that frame should be given more preference than the frame without person. To achieve the above mentioned objective, we will make use of the probability value associated with the bounding box around the person in the frame. As mentioned above, we have a set S containing the score values of each and every frame of set F of any video. For selection of frames from set F, we will put some threshold value to be achieved.

$$fr = argmax\{(C_{f1}(\varepsilon_1\,, \varepsilon_2\, ..), C_{f2}(\varepsilon_1, \varepsilon_2 ..) \cup .. C_{fn}(\varepsilon_1, \varepsilon_2 .)\}$$

$$C_{fi} = \varepsilon_1 \cup \varepsilon_2 \cup .. \varepsilon_n$$

Where, f_r = Key-frame, argmax() = selection of frame with maximum content value, C_{fj} = content measure for each frame ε_i = probabilistic scores belonging to the each object class present in the frame.

In our approach, the frame with the maximum content change function value over a shot, that also crosses a minimum threshold value, is chosen as the keyframe. The algorithm is as follows:

3.2. Sufficient Content Change

Sufficient content change is based on sequential comparison of frames with a previously chosen keyframe[31]. A new keyframe is chosen if it sufficiently differs from the previous keyframe in terms of content. The equation for sufficient content change is a follows:

$$fr = abs\{C(f_{r0}, fi)\} > \ \varepsilon, i < N_f$$

$$fr + 1 = abs\{C(fr, fi + 1)\} > \ \varepsilon, i < N_f$$

$$\mathrm{C}(f_1, f_2) = \mathrm{P}f_1 - \mathrm{P}f_2$$

Where, C = content change function, f_{k0} = base key-frame, f_i = input frame, fr= key-frame, N_f = number of frames, ε = threshold, $\mathrm{P}_{f1,\,2}$ = probabilistic scores of the frames

```
Algorithm 1 Maximum Frame Coverage
  Step 1: Take a video input
  Step 2: Convert the video into frames
  Step 3: Perform Object Detection
  Step 4: Define number of key-frames to be extracted
  Step 5: Create shots from the frames
  Step 6: Find frame with threshold $\varepsilon$ such that all other frames are covered
  Step 7: Extract that frame as key-frame
  Step 8: Combine all key-frames to generate output video
```

```
Algorithm 2 Sufficient Content Change Step 1: Take a video input
  Step 2: Convert the video into frames
  Step 3: Perform Object Detection
  Step 4: Select first frame as first key-frame
  Step 5: if (Content-Change Function > $\varepsilon$) select it as key-frame
  Step 6: For next frame, calculate difference of score w.r.t. last selected key-frame
  Step 7: Repeat for all frames in the video
  Step 8: Combine all key-frames to generate output video
```

3.3. Clustering

This method considers frames as points in the feature space, with the idea that illustrative points of clusters processed in this space can be utilized as representative frames for the whole video sequence. It does not rely on any explicit modelling while using generic approaches created for data clustering. Based on clip-by-clip or shot-by-shot clustering can be done on a, and the four processes below are frequently used [32]. Preprocessing the data, Clustering the data, Filtering clusters, and collecting the representative points of each cluster are the stages that distinguish existing approaches. Clustering appears to be a wanted method for keyframe extraction, owing to its widespread use in data analysis.

However, accurate extraction of semantic significant clusters in video data is difficult due to large intra-class visual variance and low intra-class visual variance. Existing keyframe extraction research frequently fails to thoroughly analyse the clustering process's outcomes. Furthermore, when we wish to preserve the temporal evolution of the video sequence from retrieved keyframes, the clustering-based solution is often not suitable [33].

3.4. Curve Simplification

Methods based on clustering are tied to the approach to curve simplification. The distinction is that the curve's points aren't always evenly spaced by the frame index, and there's no need for explicit error modelling between the final curve and the original frame trajectory [34].

In curve simplification depicted in Algorithm 3, the frames of a shot are mapped onto a feature space. The points are connected in temporal order to generate a trajectory curve. The curve is traversed to find a set of coordinates that, when plotted into a curve, retain the shape of the original curve. Various curve simplification algorithms have been proposed, such as binary curve splitting algorithm [35]. In our approach, we use the Ramer-Douglas-Peucker algorithm [37] for decreasing the points of the curve. The algorithm is as follows:

Algorithm 3 Curve Simplification

```
Step 1: Take a video input
Step 2: Convert the video into frames
Step 3: Perform Object Detection
Step 4: Calculate probabilistic score as $Score_i$
Step 5: Plot curve for $i$ vs $Score_i$ , where $i$ stands for the frame index
Step 6: Apply RPD Algorithm to get index of key-frames
Step 7: Repeat for all shots in the video
Step 8: Extract those indexed as key-frames
Step 9: Combine all key-frames to generate output video
```

3.5. Minimum Correlation

The problem of rate constrained keyframe extraction is frequently addressed by techniques in this class, which work on the assumption that the keyframe set's components should have minimum correlation [38]. It seeks to select frames that are unlike each other. Although the minimum frame correlation criterion in the keyframe set ensures a low level of redundancy, it suffers greatly from outliers.

The presence of the edge is contingent on a set of constraints that implement the keyframe set's temporal ordering as well as some overlapping between two consecutive keyframes, which is required for predicting the unfolding content in between. This method allows the initial and last frames to be included in the keyframe set, making the optimal keyframe set the shortest path from the first vertex or frame to the last [19]. The equation of minimum correlation is as follows:

$$f_{k1}, f_{k2} = argmin\{\mathrm{C}_1, \mathrm{C}_2 \dots \mathrm{C}_\mathrm{n}\}$$

$$C_i = Corr(f_i, f_{i+1})$$

Where, f_{k1}, f_{k2} = Key-frames, argmin() = minimum argument function checked on every frame combination in the shot, $Corr()$ = cross-correlation function on f_{r1} and f_{ri+1} frames.

```
Algorithm 4  Minimum Correlation
Step 1: Take a video input
Step 2: Convert the video into frames
Step 3: Calculate Cross-Correlation between the frames
Step 4: Estimate number of frames per shot
Step 5: Estimate minimum correlation between two frames in a shot
Step 6: Extract those frames as key-frames
Step 7: Repeat for all shots in the video
Step 8: Combine all key-frames to generate output video
```

3.6. MULTI VIEW SUMMARIZATION

In two key aspects, multi-view video summarizing differs from single-view video summary. First, despite the enormous difficulty of dealing with multi-view data, there is a structure to it. There are numerous correlations in the data due to the placements and fields of view of the cameras. So, correlations with content as well as discrepancies between different videos need to be correctly modelled for obtaining an informative summary. Second, these videos are captured with distinct angles of view and depth of view. Fields, resulting in a number of unaligned videos, for the same scenery. As a result, describing these videos is difficult due to differences in lighting, position, and view point, as well as synchronization concerns. The best methods for extracting summaries from single-view videos aren't always the most effective. Multi-view videos were summarized by a group of representatives [2][39].

Consider a set of K different videos captured from different cameras, K different frame sets will be generated. The main objective function of achieving a multi-view summarized video is achieved by the methodology proposed here. First of all, keyframes will be extracted for all the K different views of a video. Thus K different keyframe sets will be generated, containing keyframes from n different views. After the keyframes extraction from all the K different views, all the keyframes would be embedded in the same set.

$$V = \{KF11, KF12,..., KF1m, KF21, KF22 ,..., KF2m, KFn1, KFn2, ..., KFnm\}$$

Frames embedded in the same set would be sorted on the basis of temporal and similarity. Temporal order of the frames would be maintained and if the same frame number from more than 1 view is present, then only a single frame of that temporal order would be selected. This would be carried out by the comparison made on the basis of probability value of person in that frame [17]. The frame with highest probability is kept in the final set and all the other frames are eliminated. The correlation keyframe extraction technique would be applied finally to remove the similar frames. Finally, after appending all the remaining frames in the set with the help of video processing tools, a summarized video will be generated.

Algorithm 5 depicts systematic steps for generating intra-view and inter-view video summaries.

Algorithm 5 Video Summarization

```
Require: Two or more than two videos of an overlapping view
Step 1: Initialization
Step 2: for i = 1 to N do
        ConVideointoFrames(View_i,Path)
        end for
Step 3: based on object consideration frame importance is calculated by
        $f_r = argmin\{C(f_t, f_i) > \varepsilon, i < n\}$
        $f_{rj+1} = argmin\{C(f_{trj}, f_i) > \varepsilon, i < n\}$
        where, C = Content change function, $f_t$ = threshold frame, $f_i$ = input frame, n = number of
        frames, $\varepsilon$ = threshold. The content change function for this method is calculated on the basis of
        object detection algorithms, as discussed in the previous section.
Step 5: Video Similarity Graph (VSG) is generated
Step 6: Spatio-Temporal Graph from VSG is generated
Step 7: Various keyframes selection methods are implemented to generate the multi view video
```

4. Experimental Results and Discussion

We give a variety of experiments and comparisons in this section to demonstrate the usefulness and efficiency of our proposed algorithm in summarising multi-view videos.

4.1. Datasets

Dataset we have used for multi view summarization is the office dataset. Office dataset consists of 3 views capturing the same office room. The dataset we have used is a standard dataset being used for multi-view video summarization [5]. We have used office dataset for the comparison of our result.

4.2. Evaluation Techniques

In order to develop the area, the effectiveness and/or effectiveness of a novel solution to a specific problem must be tested, against existing methods. However, in video summarization research, there is a major absence of a standard evaluation framework, resulting in each work having its intrinsic evaluation method, which often lacks a comparison of performance with previous techniques. This is somewhat due to the lack of an objective ground-truth, which makes judging the correctness of a video abstract more difficult than in other research areas such as object identification and recognition [32]. Even for humans, deciding if one summarized video is better than another is tough, and to make matters difficult, the generated summary perspective are sometimes application-dependent. Conciseness, coverage, context, and coherence are the four qualities. For a keyframe set, the first two characteristics are also desirable [14].

For the better evaluation, our team members had selected out the keyframes according to their knowledge, selection was done on the basis of number of people in the frame, movement best captured from the 3 views. The selected set of frames was then compared with the final summarized video frames.

4.2.1. Precision and Recall

They are estimated over test series. In the field of image classification, information retrieval and video segmentation these two are usually employed.

$$Precision = \frac{No. of\ images\ classified\ accurately}{Total\ no. of\ images\ classified}$$

$$Recall = \frac{Number\ of\ images\ calssified\ accurately}{Total\ no. of\ images\ in\ the\ database}$$

4.2.2. Compression Ratio and F-Measure

The Compression Ratio (CR) is a measurement of the shot's compactness as a result of keyframe selections, and it varies depending on the amount of keyframes utilised. The following equation is used to calculate the compression ratio:

$$CR = \frac{Number\ of\ Frames\ in\ Output\ Video}{Total\ Number\ of Frames\ In\ Video}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.3. Result Discussion

A comparison of various methods is the most important part in any research oriented project. For our multi-view video summarization, we carried out comparison between the top 4 performing key frame extraction methods. Our comparison is based on the number of frames selected by us which can be treated as keyframe and number of frames extracted by the key frame extraction methods. Then the number of intersection frames are calculated. After that, browsing and retrieval are used to determine the accuracy of the particular key frame extraction method. Video abstracts that are concise and intelligently generated will make it easier for users to access enormous amounts of video content in an effective and efficient manner. Researchers have recently shown a strong interest in video abstraction, and as a result, a variety of algorithms and techniques have been proposed. We conducted a detailed assessment and analysis of the research in two major types of video abstraction: the keyframe set and the skim. For calculating the accuracy, the number of intersection frames are compared to the number of frames extracted by algorithm.

Table 1, 2, and 3 shows the video summary generated by different views for the same area with different angle. Table 4 shows the multi-view summary generated by using different key-frame extraction methods. It is evident from table 3 that for close room scenario maximum frame coverage method gives higher accuracy for inter-view stage of multi-view video summarization.

	Frames Selected by User	Frames Selected by Algorithm	Intersections	Accuracy

			of Frames	
Maximum Frame Coverage	100	110	96	87
Sufficient Content Change	100	62	60	97
Curve Simplification	100	80	76	95
Minimum Co-relation	100	30	28	93

Table 1: View 1 Summary

	Frames Selected by User	Frames Selected by Algorithm	Intersections of Frames	Accuracy
Maximum Frame Coverage	100	87	84	97
Sufficient Content Change	100	64	60	94
Curve Simplification	100	78	76	97
Minimum Co-relation	100	30	26	87

Table 2: View 2 Summary

	Frames Selected by User	Frames Selected by Algorithm	Intersections of Frames	Accuracy
Maximum Frame Coverage	100	96	89	93
Sufficient Content Change	100	67	61	91
Curve Simplification	100	87	81	93
Minimum Co-relation	100	30	25	83

Table 3: View 3 Summary

	Frames Selected by User	Frames Selected by Algorithm	Intersections of Frames	Accuracy
Maximum Frame Coverage	100	105	95	90
Sufficient Content Change	100	45	40	89
Curve Simplification	100	78	68	87

Minimum Co-relation	100	30	23	77

Table 4: Multi-view Summary

5. Conclusion

The work discussed in the paper contains details of keyframe extraction technique and parallel processing algorithm. The keyframe extraction system developed by referring all mentioned paper for identifying which key-frame extraction techniques perform best for closed room scenario. It generates a summary video from multiple videos. For closed room scenario maximum frame coverage method gives higher accuracy for inter-view stage of multi-view video summarization

References

[1] T. Huang, Surveillance video: The biggest big data, Computing Now 7 (2) (2014) 82–91.

[2] S. M. Metev, V. P. Veiko, Laser-assisted microtechnology, Vol. 19, Springer Science & Business Media, 2013.

[3] J. Breckling, The analysis of directional time series: applications to wind speed and direction, Vol. 61, Springer Science & Business Media, 2012.

[4] B. T. Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM transactions on multimedia computing, communications, and applications (TOMM) 3 (1) (2007) 3–es.

[5] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering., in: Nips, Vol. 14, 2001, pp. 585–591.

[6] Y. Cong, J. Yuan, J. Luo, Towards scalable summarization of consumer videos via sparse dictionary selection, IEEE Transactions on Multimedia 14 (1) (2011) 66–75.

[7] S. Feng, Z. Lei, D. Yi, S. Z. Li, Online content-aware video condensation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2082–2087.

[8] J. Peng, Q. Xiao-Lin, Keyframe-based video summary using visual attention clues, IEEE MultiMedia 17 (02) (2010) 64–73.

[9] J. C. Caicedo, S. Lazebnik, Active object localization with deep reinforcement learning, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2488–2496.

[10] C. Chennubhotla, A. Jepson, Sparse coding in practice, in: Proc. of the Second Int. Workshop on Statistical and Computational Theories of Vision, 2001.

[11] W.-S. Chu, Y. Song, A. Jaimes, Video co-summarization: Video summarization by visual co-occurrence, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3584–3592.

[12] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, S. Li, Flickr distance, in: Proceedings of the 16th ACM international conference on Multimedia, 2008, pp. 31–40.

[13] K. Muhammad, T. Hussain, S. W. Baik, Efficient cnn based summarization of surveillance videos for resource-constrained devices, Pattern Recognition Letters 130 (2020) 370–375.

[14] Y. Yao, Three-way decisions with probabilistic rough sets, Information sciences 180 (3) (2010) 341–353.

[15] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, F. A. Shah, Video summarization: techniques and classification, in: International Conference on Computer Vision and Graphics, Springer, 2012, pp. 1–13.

[16] L. Herranz, J. M. Mart´ınez, An efficient summarization algorithm based on clustering and bitstream extraction, in: 2009 IEEE International Conference on Multimedia and Expo, IEEE, 2009, pp. 654–657.

[17] S. D. Thepade, A. A. Tonge, An optimized key frame extraction for detection of near duplicates in content based video retrieval, in: 2014 International Conference on Communication and Signal Processing, IEEE, 2014, pp. 1087–1091.

[18] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: European conference on computer vision, Springer, 2014, pp. 505–520.

[19] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, A. de Albuquerque Araujo,´ Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters 32 (1) (2011) 56–68.

[20] J. Almeida, N. J. Leite, R. d. S. Torres, Vison: Video summarization for online applications, Pattern Recognition Letters 33 (4) (2012) 397–409.

[21] E. Elhamifar, R. Vidal, Sparse subspace clustering: Algorithm, theory, and applications, IEEE transactions on pattern analysis and machine intelligence 35 (11) (2013) 2765–2781.

[22] R. Panda, A. Dasy, A. K. Roy-Chowdhury, Video summarization in a multi-view camera network, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2971–2976.

[23] S. Gao, I. W.-H. Tsang, L.-T. Chia, P. Zhao, Local features are not lonely– laplacian sparse coding for image classification, in: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, 2010, pp. 3555– 3561.

[24] J. C. S. Yu, M. S. Kankanhalli, P. Mulhen, Semantic video summarization in compressed domain mpeg video, in: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), Vol. 3, IEEE, 2003, pp. III–329.

[25] D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, in: 2007 IEEE 11th international conference on computer vision, IEEE, 2007, pp. 1–8.

[26] E. Elhamifar, G. Sapiro, R. Vidal, See all by looking at a few: Sparse modeling for finding representative objects, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 1600–1607.

[27] C.-W. Ngo, T.-C. Pong, R. T. Chin, Video partitioning by temporal slice coherency, IEEE Transactions on Circuits and Systems for Video Technology 11 (8) (2001) 941–953.

[28] F. Dornaika, I. K. Aldine, Decremental sparse modeling representative selection for prototype selection, Pattern Recognition 48 (11) (2015) 3714–3727.

[29] P. Chiu, A. Girgensohn, Q. Liu, Stained-glass visualization for highly condensed video summaries, in: 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), Vol. 3, IEEE, 2004, pp. 2059–2062.

[30] Z. ¯ Cernekova, C. Nikou, I. Pitas, Entropy metrics used for video summarization,´ in: Proceedings of the 18th spring conference on Computer graphics, 2002, pp. 73–82.

[31] R. Zabih, J. Miller, K. Mai, A feature-based algorithm for detecting and classifying scene breaks, in: Proceedings of the third ACM international conference on Multimedia, 1995, pp. 189–200.

[32] I. S. Lim, D. Thalmann, Key-posture extraction out of human motion data, in: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 2, IEEE, 2001, pp. 1167–1169.

[33] X. Zeng, W. Li, X. Zhang, B. Xu, et al., Key-frame extraction using dominantset clustering, in: 2008 IEEE international conference on multimedia and expo, IEEE, 2008, pp. 1285–1288.

[34] C. Cotsaces, N. Nikolaidis, I. Pitas, Video shot detection and condensed representation. a review, IEEE signal processing magazine 23 (2) (2006) 28–37.

[35] A. Nasreen, G. Shobha, Key frame extraction from videos-a survey, International Journal of Computer Science & Communication Networks 3 (3) (2013) 194.

[36] D. G. Lowe, Three-dimensional object recognition from single two-dimensional images, Artificial intelligence 31 (3) (1987) 355–395.

[37] W. Mokrzycki, M. Samko, Canny edge detection algorithm modification, in: International Conference on Computer Vision and Graphics, Springer, 2012, pp. 533–540.

[38] G. Guan, Z. Wang, S. Mei, M. Ott, M. He, D. D. Feng, A top-down approach for video summarization, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 11 (1) (2014) 1–21.

[39] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, Z.-H. Zhou, Multi-view video summarization, IEEE Transactions on Multimedia 12 (7) (2010) 717–729.

[40] S. K. Kuanar, K. B. Ranga, A. S. Chowdhury, Multi-view video summarization using bipartite matching constrained optimum-path forest clustering, IEEE Transactions on Multimedia 17 (8) (2015) 1166–1173.

[41] V. Parikh, J. Mehta, S. Shah, and P. Sharma, "Comparative analysis of keyframe extraction techniques for video summarization," Recent Advances in Computer Science and Communications, vol. 14, 2021.

[42] V. Parikh, P. Sharma, V. Shah, and V. Ukani, "Optimal camera placement for multimodal video summarization," in *International Conference on Futuristic Trends in Network and Communication Technologies*, pp. 123–134, Springer, 2018.