# Feature Extraction and Clustering for Static Video Summarization

**Yunyun Sun**
Nanjing University of Posts and Telecommunications
**Peng Li** ( lipeng@njupt.edu.cn )
Nanjing University of Posts and Telecommunications
**Yutong Liu**
Nanjing University of Posts and Telecommunications
**Zhaohui Jiang**
Nanjing University of Posts and Telecommunications

**Research**

**Keywords:** Cluster, feature data, threshold, optimization, key frame

**Posted Date:** March 30th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-344569/v1

RESEARCH

# Feature extraction and clustering for static video summarization

Yunyun Sun[1], Peng Li[2,3*], Yutong Liu[2] and Zhaohui Jiang[4]

*Correspondence:
1278369168@qq.com
[1]School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
Full list of author information is available at the end of the article

**Abstract**

Numerous limitations of shot based and content based key frame extraction approaches have encouraged the development of cluster based methods. This work provides OTMW, Optimal Threshold and Maximum Weight clustering method, as a novel cluster based key frame extraction method. The video feature dataset is constructed by computing the color, texture and information complexity features of frame images. An optimization function is developed to compute the optimal clustering threshold $\varrho$. It is constrained by fidelity and ratio measure parameters. We turn to an empirical study on the proposed method in multi-type video key frame extraction tasks and compare it with popular cluster based methods including Mean-shift, DBSCAN, GMM and K-means. OTWM method achieves an average fidelity and ratio of 96.12 and 97.13, respectively. Experimental results demonstrate that OTMW can bring higher fidelity and ratio performance, while still maintaining a competitive performance over other cluster based methods. Overall, the proposed method can accurately extract key frames from multi-type videos.

**Keywords:** Cluster; feature data; threshold; optimization; key frame

## 1 Introduction

The concept of video summarization allows the users to browse the video data in a friendly and quick way without losing key information. Since the 1990s, video summarization technology has been studied extensively because of its simplicity and usefulness[1].

Static video summarization is the collection of representative frames (key frames) which are extracted from the original video sequence. Key frame extraction methods can be divided into three categories: shot based, content based and cluster based. Currently, some shot based techniques are developed in the area of computer vision and image processing. Huang C. extracts representative frames from each shot by computing the frame image difference in saliency and edge map features[2]. Mehmood I. analyzes the difference between frame images in a shot by modeling an auditory and perceptual attention features[3]. Song G. H. computes the color difference in one shot by employing the average histogram method[4]. It is common for shot based methods to segment the original video into several shots at first. However, the shot segmentation process is computationally expensive. Content based methods can avoid this problem. Rachida H. proposes MSKVS, a content based method, to measure the inter-frame distance by time and visual features. MSKVS guarantees superior performance over other content based methods[5]. Gianluigi C. conducts experiment on six new and sport competition videos by employing his

content based method. Experimental results demonstrate that his method can effectively extract key frames[6]. Generally, these content based methods analyze the video content by extracting color, texture or motion feature. A limiting factor of content based methods is that the computational cost is incurred in the process of frame image features[7]. This limitation encourages the development of cluster based methods. Cluster based techniques work by clustering together the similar frames and extracting one representative frame of each class. They avoid shot segmentation error of shot based methods, decrease inter-frame difference analysis frequency of content based methods, and can be well-suited to key frame extraction in related fields[8, 9, 10].

In this paper, a novel cluster based key frame extraction approach is proposed. The benefits of this approach are as follows:

- A video content analysis method is proposed to fully characterize the video information by using three visual features: color, texture and information complexity.
- We develop a threshold optimization function to alleviate the task of manual choosing a cluster threshold.
- We utilize the fusion of the frame density, inter-cluster distance and intra-cluster distance to filter the key frame candidates and employ the max weight factor parameter to further refine key frame candidates.

The rest of this paper is organized as follows. Section 2 describes the details of cluster based methods. In section 3, we present the implementation of feature extraction method. Section 4 provides a detailed description of the proposed cluster based key frame extraction method. In section 5, we explore the performance of the proposed approach. Finally, the major work is discussed and wrapped up in Section 6.

## 2 Related work

The basic cluster based methods can be categorized into two: automatic and semi-automatic cluster based methods. In general, semi-automatic cluster based method requires manual determination the initial cluster centers and the number of clusters. This method is widely used in early time, but is not applicable to current abundant videos. Among the automatic cluster schemes, Kuanar uses Delaunay method and analyzes video contents by computing color and texture features[8]. Liu and his colleagues calculate the initial cluster centers by employing hierarchical method[11]. Another work uses the spectral algorithm to cluster the color histogram of frame images[10]. In artificial intelligence research community, automatic cluster based methods have a bright application prospect[9]. The core idea behind such automatic cluster based methods is to set a favorable threshold. In literature, researchers usually receive the threshold by defining formula or by setting fixed value. For example, Kuanar S. K. computes the cluster threshold by employing formula $2(1-\varepsilon)$[8]. Jeong D. J. selects 0.0001 as the cluster threshold[10]. The other researchers compute cluster threshold by a self-defined formula[9, 11]. These methods habitually neglect the mutation characteristic of individual in videos. In representative frame extraction, some cluster based techniques take the cluster centers or centroids as the representative frames of each class. In [8], the author selects the frames which are

closest to the cluster centroid as the representative frames. In [11, 12], the authors directly extract cluster centers as the representative frames. These methods evaluate the representative of the frames in one cluster by a single image feature. However, a single image feature can not being able to fully characterize the frame content and complexity. In an effort to improve this problem, researchers provide some variants to compute the representative of the frames by extracting multiple features including entropy, motion information or region of interest[13, 14, 15, 16, 17].

## 3 Methods

In this section, we provide a detailed description of the proposed OTMW method which includes feature extraction and key frame extraction. At first, the color, texture, and information complexity features are computed to express video content. Then, an optimization function is developed to compute the optimal clustering threshold. Next, the frame density, inter-distance and intra-distance are computed and fused as the clustering weight factor. Finally, a Max Weigh method is proposed to extract the cluster representative frame. The proposed approach is summarized in figure 1.

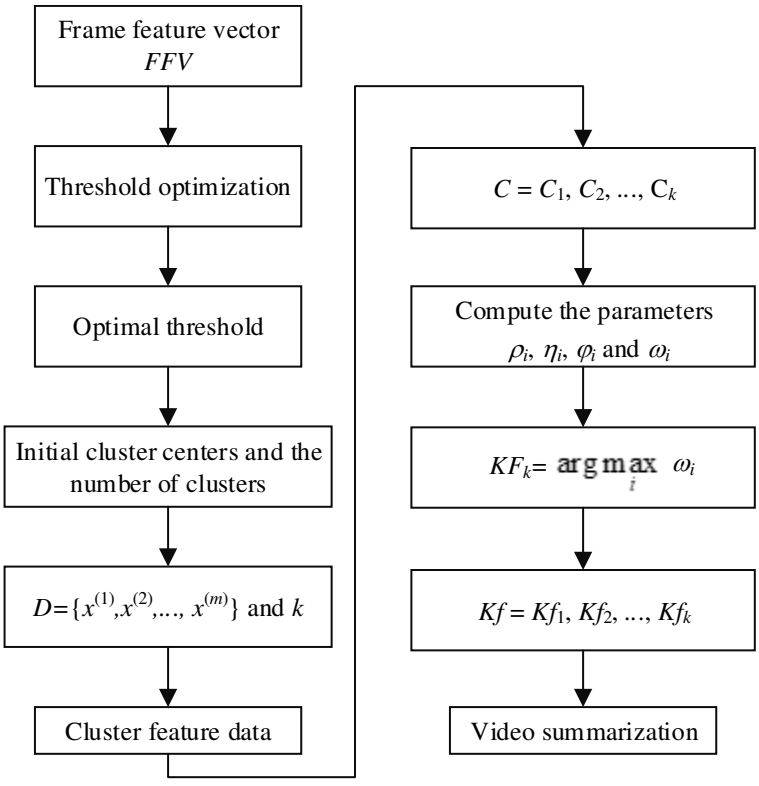


Figure 1: Framework of the proposed cluster based method.

### 3.1 Feature extraction

In this section, we describe the proposed video content analysis method which distinguishes frames by computing feature data[18]. We extract the color, texture

and information complexity features to discriminate different frame images. The video frame feature data is construct by

$$FFV = [CTE] \tag{1}$$

where $C$, $T$ and $E$ represent the color, texture and information complexity, respectively.

*3.1.1 Color feature*

We take the color feature as one feature to characterize the difference of frame images. We compute the first color moment, second color moment and third color moment in H, S, and V channels to construct the color feature data vectors of frame images. The first color moment reflects the brightness difference, which is calculated by

$$C_m = \frac{1}{w \times h} \sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q) \tag{2}$$

where the parameters $w$ and $h$ are the pixel width and height, $f_i(x_p, y_q)$ is the pixel value in position $(x_p, y_q)$, and $1 \leqslant p \leqslant w, 1 \leqslant q \leqslant h$. The second color moment reflects the color distribution range, which is compute by

$$C_v = (\frac{1}{w \times h} \sum_{p=1}^{w} \sum_{q=1}^{h} (f_i(x_p, y_q) - C_m)^2)^{\frac{1}{2}} \tag{3}$$

The third color moment represents the color distribution symmetry, which is computed by

$$C_s = (\frac{1}{w \times h} \sum_{p=1}^{w} \sum_{q=1}^{h} (f_i(x_p, y_q) - C_v)^3)^{\frac{1}{3}} \tag{4}$$

where $C_m$, $C_v$ and $C_s$ include first moment mean, second moment variance and third moment slope three parameters, respectively.

*3.1.2 Texture feature*

We take the texture feature as another feature to characterize the difference of frame images in image surface structural organization information[19]. We compute the mean of angular second moment, contrast, correlation and homogeneity texture features in 0, 45, 90 and 135 directions to construct video frame texture feature data vectors. The angular second moment characterizes the thickness and gray distribution uniformity of images, and it can be calculated by

$$T_A = \frac{1}{w \times h} \sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q)^2 \tag{5}$$

The contrast characterizes the groove depth and clarity of images, it can be calculated by

$$T_{Con} = \frac{1}{w \times h} \sum_{p=1}^{w} \sum_{q=1}^{h} (p-q)^2 f_i(x_p, y_q) \tag{6}$$

The correlation characterizes the local gray similarity in row or column direction, it can be calculated by

$$T_{Cor} = \frac{\sum_{p=1}^{w} \sum_{q=1}^{h} (x_p - \sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q) \times p) \times (y_p - \sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q) \times q) \times f_i(x_p, y_q)^2}{\sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q) \times (x_p - \sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q) \times p)^2} \tag{7}$$

The homogenization characterizes the local gray level uniformity of images, it can be calculated by

$$T_H = \sum_{p=1}^{w} \sum_{q=1}^{h} f_i(x_p, y_q) \times \frac{1}{1 + (p-q)^2} \tag{8}$$

*3.1.3 Information complexity feature*

We take the information complexity as the last feature to characterize the difference of frame images in aggregated and spatial feature. Information entropy measures the information complexity of images from holistic perspective, which is proposed by Shannon[20]. The bigger information entropy means the bigger internal non-uniformity degree. The two-dimensional information entropy $Ef_i$ can be calculated by

$$E = -\frac{1}{w \times h} Cf \times \log_2 Cf \tag{9}$$

where $Cf$ is the occurrence probability of each gray level in $i$-th frame image.

## 3.2 Clustering for key frame extraction

In this section, we describe the proposed cluster based key frame extraction method, which develops a new optimization function to compute the optimal threshold.

*3.2.1 Threshold optimization*

We narrow the search interval of optimal threshold by computing the function values of trial points. In cluster based method, the fidelity[6] and ratio[5] are negatively and positively correlated with the threshold, respectively. Therefore, we infer that the quality of key frames is optimal when fidelity and ratio are infinitely close. We introduce a new parameter FR to characterize this relationship and to obtain

the optimal key frames. The distance between frame $x^{(i)}$ and frame $x^{(j)}$ is compute by

$$d_{ij} = \|x^{(i)} - x^{(j)}\|_2 = \sqrt{\sum_{u=1}^{n} |x_u^{(i)} - x_u^{(j)}|^2} \tag{10}$$

where $(i, j) \in [1, 2, ..., m]$. The average distance is compute by

$$d_c = \frac{2}{m(m-2)} \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} \tag{11}$$

The threshold is defined as

$$t = d_c \pm \varepsilon \times std_{ij} \tag{12}$$

where $\varepsilon$ is a variable factor, $std_{ij}$ is the standard deviation of $d_{ij}$. We define the new parameter FR as

$$FR(t) = \frac{fidelity(t)}{ratio(t)} \tag{13}$$

The threshold optimization function is defined as

$$f(t) = \mid FR(t) - 1 \mid \tag{14}$$

We compute the optimal threshold by following steps. The new parameters $a = d_c - 3 \times std_{ij}$ and $b = d_c + 3 \times std_{ij}$.

Step1: We compute $f(a)$, $f(b)$ and $f(c)$, where $c = a + 0.618 * (b - a)$.

Step2: If $f(a) = f(b) = f(c)$, turn to Step4, else turn to Step3.

Step3: If $f(c) < 0$, we change $c$ to increase fidelity and decrease ratio, $b = c$; Else we change $c$ to decrease fidelity and increase ratio, $a = c$. Then, return to Step1.

Step4: Here $c = a + 0.382 * (b - a)$ and return to Step1.

Step5: If $f(c) = f(a) = f(b)$ and $(b - a) \leq 0.001$ in three successive computation, turn to Step6. Else return to Step4.

Step6: We compute the optimal threshold $\varrho$ by $\varrho = \frac{a+b}{2}$.

*3.2.2 Initial cluster centers and the number of clusters*

In this section, we compute the initial cluster centers and the number of clusters by clustering the data of $FFV$. The process is shown in figure 2. The pseudo code of the proposed approach is given in Algorithm 1. The density $\rho_i$ is calculated by

$$\rho_i = \sum_j e^{-(\frac{d_{ij}}{d_c})^2} \tag{15}$$

The distance between $j$ and $i$ is defined as

$$\eta_i = dist(x^{(i)}, x^{(j)}), i \in D, j \neq i \tag{16}$$

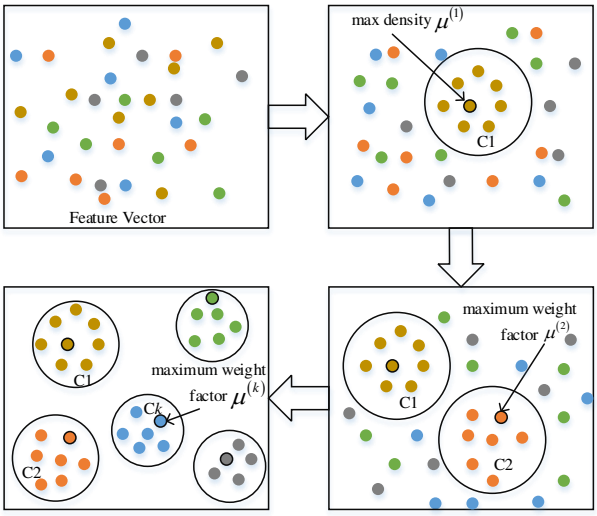


Figure 2: The computation of initial cluster centers and the number of clusters

where $i$ and $j$ are in the same class. The distance between $i$ and $j$ is defined as

$$\varphi_i = dist(x^{(i)}, x^{(j)}), i \in D, j \in \mu^k \tag{17}$$

here $j$ represents the cluster center, and $j$ and $i$ are in different classes. The weight factor $\omega_i$ is defined as

$$\omega_i = \Pi(\rho_i \times \eta_i^{-1} \times \varphi_i) \tag{18}$$

*3.2.3 Key frame extraction*

In this section, we classify frame images into $k$ clusters $C = \{C_1, C_2, ..., C_k\}$ and extract representative frames from this clusters. The error square sum criterion function is used as the criterion function. The frame images are classified into different clusters by employing Algorithm 2. We calculate the parameters $\rho_i$, $\eta_i$, $\varphi_i$ and $\omega_i$ of clusters $C = \{C_1, C_2, ..., C_k\}$ and extract the representative frame of cluster $C_k$ by $KF_k = \arg\max_i \omega_i$.

## 4 Results and discussion

In this section, we turn to an empirical study on the proposed OTMW method in key frame extraction tasks and compare it with cluster based methods including Mean-shift[21], DBSCAN[22], GMM[23] and K-means[24]. We abbreviate these cluster based methods as Ms, DB, GM, Km and my, respectively. We report improved performance across open video dataset. We conduct a set of experiments by using surveillance, documentary, lecture on TV and phone recording four different video datasets. These videos are publicly shared on https://open-video.org/. In this section, we take $Hcil2000\_01$ video as an example to report the performance of OTMW method. Here $Hcil2000\_01$ is a random video of open video dataset.

**Algorithm 1** Compute the initial cluster centers and the number of clusters

**Input:** $FFV = \{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$

**Output:** the initial cluster centers $\{\mu^{(1)}, \mu^{(2)}, ..., \mu^{(k)}\}$ and the number of clusters $k$

compute the distance $d_c$;
**for** each sample $i \in D$ **do**
    compute the density $\rho_i$;
**end for**
**while** $D \neq \varnothing$ **do**
    set the frame with maximum $\rho_i$ as the initial cluster center $\mu^{(1)}$;
    **if** $d(\mu^{(1)}, i) \leq t$ **then**
        $i \in C_1$, remove $i$ from $D$;
    **end if**
    **for** each sample $i \in D \cup (C_i \notin D)$ **do**
        compute $\eta_i, \varphi_i, \omega_i$;
        cluster center $\mu^{(i)} = \arg\min_i \omega_i$
        $i \in C_i$, remove $i$ from $D$;
    **end for**
**end while**

### 4.1 Optimization of threshold

The parameters of $Hcil2000\_01$ video in threshold optimization are shown in table 1. In $Hcil2000\_01$ video, the average distance $d_c$ and standard deviation $std_{ij}$ of $d_{ij}$ are 2.3107 and 0.2259, respectively. Therefore, the parameters $a = 1.6442, b = 2.9993$. The variable interval of parameter $c$ is [1.6442,2.9993]. The parameter $c$ is computed by $c = a + 0.618 * (b - a) = 2.481652$. In sixth iteration, the $f(a) = f(b) = f(c) = 0.0063$. The calculation of parameter $c$ is changed to $c = a + 0.382 * (b - a)$ in subsequent iterations. As shown in table1, the value of $f(a) = f(b) = f(c) = 0.0063$ are not change in the next two iterations. However, $b - a = 0.122156 > 0.001$. Finally, $b - a = 0.000509 < 0.001$ in 13th iteration. Therefore, the optimal threshold $\varrho = \frac{a+b}{2} = 1.644709$.

### 4.2 Extraction of key frame

The key frames are extracted by employing OTMW method across open video dataset. The key frames of $Hcil2000\_01$ video are shown in figure 3. The fidelity and ratio results are shown in table 2. The fidelity measures of different videos are changed from 93 to 98 with an average of 96.12. The ratio measures are changed from 95 to 98 with an average of 97.13. The key frames are consistent with artificial judgment.

### 4.3 Comparisons between OTMW and other cluster based methods

We compare OTMW with popular cluster based algorithm in term of the fidelity and ratio measure performance. In experimental, the number of clusters of semi-automatic cluster based methods are same as OTMW method. The results of fidelity and ratio are shown in figure 4 and table 3. Mean-shift cluster based method with

**Algorithm 2** Cluster the frame feature value

**Input:** frame feature value $FFV = \{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$, the initial cluster centers $\{\mu^{(1)}, \mu^{(2)}, ..., \mu^{(k)}\}$ and the number of clusters $k$.

**Output:** $k$ clusters $C = \{C_1, C_2, ..., C_k\}$.

Let $C_i = \varnothing (1 \leq i \leq k)$;
**repeat**
  **for** $j = 1, 2, ..., m$ **do**
    compute the distance between the sample $x^{(j)}$ and the cluster center $\mu^{(i)} (1 \leq i \leq k)$;
    compute the $\lambda_j = argmind_{ji}$;
    divide the sample $x^{(j)}$ into the nearest cluster $C_{\lambda j} = C_{\lambda j} \cup x^{(j)}$;
  **end for**
  **for** $i = 1, 2, ..., k$ **do**
    calculate the new cluster center $(\mu^{(i)})' = \frac{1}{|C_i|} \sum_{x \in C_i} x$;
    **if** $(\mu^{(i)})' = \mu^{(i)}$ **then**
      update the current cluster center $\mu^{(i)}$ to $(\mu^{(i)})'$;
    **else**
      keep the current mean vector unchanged;
    **end if**
  **end for**
**until** the current cluster center vectors are not updated.

Table 1: The parameters of $Hcil2000\_01$ video in threshold optimization. Here $f'(a) = 100 * f(a)$, $f'(b) = 100 * f(b)$ and $f'(c) = 100 * f(c)$.

| *iterations* | $a$ | $b$ | $c$ | *f'(a)* | *f'(b)* | *f'(c)* |
|---|---|---|---|---|---|---|
| 1 | 1.6442 | 2.9993 | 2.4817 | -0.63 | -7.6 | -6.95 |
| 2 | 1.6442 | 2.4817 | 2.1617 | -0.63 | -6.95 | -3.33 |
| 3 | 1.6442 | 2.1617 | 1.9640 | -0.63 | -3.33 | -2.11 |
| 4 | 1.6442 | 1.9640 | 1.8419 | -0.63 | -2.11 | -2.11 |
| 5 | 1.6442 | 1.8419 | 1.7664 | -0.63 | -2.11 | -0.63 |
| 6 | 1.6442 | 1.7663 | 1.7197 | -0.63 | -0.63 | -0.63 |
| 7 | 1.6442 | 1.7197 | 1.6720 | -0.63 | -0.63 | -0.63 |
| 8 | 1.6442 | 1.6720 | 1.6544 | -0.63 | -0.63 | -0.63 |
| 9 | 1.6442 | 1.6544 | 1.6479 | -0.63 | -0.63 | -0.63 |
| 10 | 1.6442 | 1.6479 | 1.6455 | -0.63 | -0.63 | -0.63 |
| 11 | 1.6442 | 1.6455 | 1.6447 | -0.63 | -0.63 | -0.63 |
| 12 | 1.6442 | 1.6447 | 1.6443 | -0.63 | -0.63 | -0.63 |
| 13 | 1.6442 | 1.6447 | 1.6445 | -0.63 | -0.63 | -0.63 |

an average fidelity of 83.75 and an average ratio of 95.59. DBSCAN cluster based method with an average fidelity of 85.75 and an average ratio of 94.79. The average fidelity of K-means and GMM cluster based methods are 85.61 and 85.38. The proposed OTMW method with an average fidelity of 96.24 and with an average ratio of 97.15. OTMW method achieves a 10.63-12.49 fidelity improvement over other cluster based methods. The fluctuations of ratio measure of different videos are shown in figure 5. OTMW method with a ratio variance of 0.73. The ratio variance of Mean-shift and DBSCAN cluster based methods are 22.11 and 11.12. They are 15 and 30 times larger than OTMW, respectively. OTMW method achieves


Figure 3: Results of the video from Open video dataset.

Table 2: The fidelity and ratio performance of videos. Here Nf represents the number of frames, Nrf represents the number of key frames.

| video name | Nf | Nrf | fidelity | ratio |
|---|---|---|---|---|
| Traffic monitoring | 120 | 5 | 95.94 | 95.83 |
| Office video_01 | 161 | 5 | 96.84 | 96.89 |
| Entertainment_01 | 151 | 5 | 97.14 | 96.69 |
| Entertainment_02 | 191 | 7 | 98.01 | 96.34 |
| Entertainment_03 | 101 | 4 | 97.79 | 96.04 |
| Office video_02 | 77 | 3 | 96.34 | 96.10 |
| Indi012 | 201 | 6 | 94.53 | 97.00 |
| UGS01_008 | 301 | 5 | 93.58 | 98.34 |
| UGS07_005 | 400 | 8 | 95.07 | 98.00 |
| UGS01_006 | 360 | 8 | 95.81 | 97.78 |
| UGS01_001 | 331 | 7 | 94.26 | 97.89 |
| Hcil2000_01 | 210 | 7 | 95.58 | 96.67 |
| Marchionini | 100 | 4 | 96.96 | 96.00 |
| RayDiessenIBM | 230 | 6 | 96.73 | 97.39 |
| Entertainment_04 | 201 | 3 | 97.63 | 98.51 |
| Daily life | 283 | 5 | 97.45 | 98.23 |

a 1.56-2.24 ratio improvement over other cluster based methods and has a small fluctuation.

**Extraction of key frames on various datasets**. To assess the performance of OTMW method, we consider key frame extraction tasks on surveillance, documentary, lecture on TV and phone recording datasets. The mean-shift cluster based method achieves an average fidelity of 82.42, 86.90,84.36 and 81.22, respectively. It achieves an average ratio of 95.76,92.73, 94.767 and 99.05. DBSCAN cluster based method achieves an average fidelity of 88.40, 84.07, 81.02 and 89.57. It achieves an average ratio of 92.19, 97.37, 97.42 and 92.13. The average fidelities of K-MEANS cluster based method are 86.39, 83.77, 85.51 and 87.62. The average ratios of GMM cluster based method are 86.39, 83.18, 85.94 and 86.45. OTMW method achieves an average fidelity of 97.07, 94.40, 95.87 and 97.54 and an average ratio of 96.43, 97.83, 97.01 and 98.37. The fidelity measures on various datasets are shown in figure 6. OTMW method achieves a 9.91-11.66 fidelity and 0.91-2.77 ratio improvement over Mean-shift, DBSCAN, GMM and K-MEANS cluster based methods.

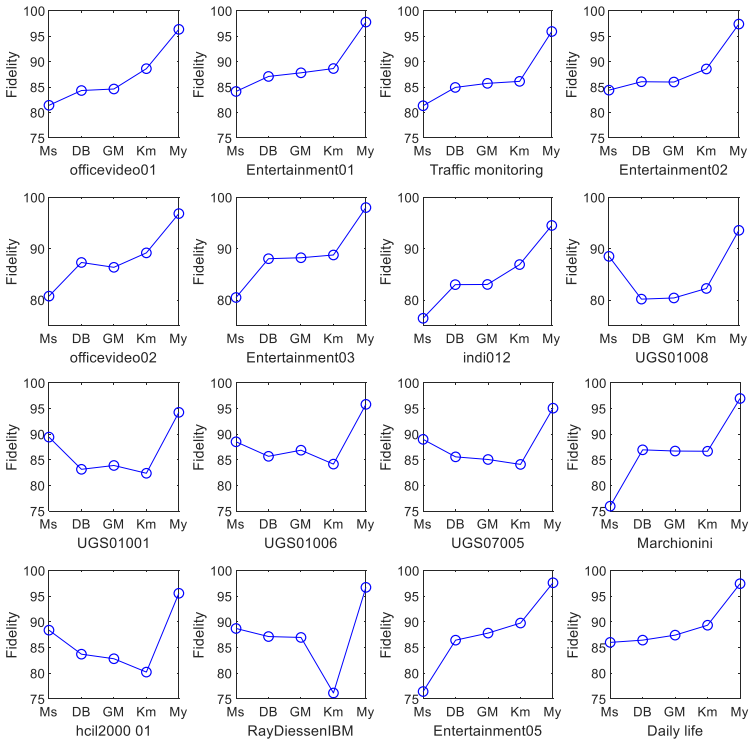


Figure 4: The fidelity measure performance of cluster based methods

## 5 Conclusions

In this paper, an innovative cluster based key frame extraction method is presented for multi-type videos. This method analyzes the video content by extracting the color, texture and information complexity features. The threshold optimization function is constrained by fidelity and ratio measures. It avoids the dependence on a fixed threshold problem in traditional cluster based method by computing the optimal threshold. The parameters density $\rho_i$, inter-distance $\eta_i$, intra-distance $\varphi_i$ and weight factor $\omega_i$ are used to compute the initial cluster centers and the number of clusters, extract representative frames from $k$ clusters. The method shows promising result on different video datasets. Meanwhile, OTMW achieves competitive and even better fidelity and ratio measure performance when compared with other cluster based methods. Overall, we found that OTMW well suited to process key frame extraction problem in the field of static video summarization.

## Appendix

**Acknowledgements**
Not applicable

**Funding**
The subject is sponsored by the National Key R&D Program of China (No. 2018YFB1003201), the National Natural Science Foundation of P. R. China (No. 61672296, No. 61602261, No. 61872196, and No. 61872194), Scientific and Technological Support Project of Jiangsu Province (No. BE2017166, and No. BE2019740), Major Natural Science Research Projects in Colleges and Universities of Jiangsu Province (No. 18KJA520008), Major Natural Science Research Projects in Colleges and Universities of Anhui Province(No.KJ2019ZD20).

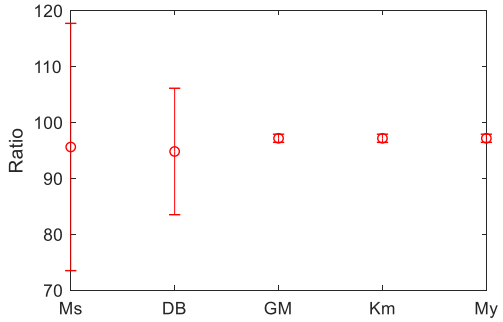


Figure 5: The ratio error-bar of cluster based methods

Table 3: The ratio measure performance of cluster based methods

| video name | Ms | DB | My |
|---|---|---|---|
| Traffic monitoring | 97.5 | 95.0 | 95.83 |
| Officevideo_01 | 98.76 | 88.20 | 96.89 |
| Entertainment_01 | 81.46 | 92.05 | 96.67 |
| Entertainment_02 | 98.43 | 93.72 | 96.34 |
| Entertainment_03 | 98.14 | 94.06 | 96.04 |
| Officevideo_02 | 98.70 | 89.61 | 96.10 |
| Indi012 | 99.50 | 94.50 | 97.02 |
| UGS01_008 | 95.26 | 98.50 | 98.34 |
| UGS07_005 | 93.75 | 98.50 | 98.00 |
| UGS01_006 | 96.11 | 97.78 | 97.78 |
| UGS01_001 | 89.43 | 98.19 | 97.89 |
| Hcil2000_01 | 90.95 | 97.14 | 96.67 |
| Marchionini | 99.00 | 96.00 | 96.96 |
| RayDiessenIBM | 94.35 | 99.13 | 97.39 |
| Entertainment_04 | 99.50 | 89.55 | 98.51 |
| Daily life | 98.58 | 94.70 | 98.23 |

**Abbreviations**

OTMW:Optimal Threshold and Maximum Weight;DBSCAN:Density-Based Spatial Clustering of Applications with Noise; GMM:Gaussian Mixed Model

**Availability of data and materials**

All data generated or analysed during this study are included in this published article.

**Ethics approval and consent to participate**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

**Consent for publication**

Not applicable

**Authors' contributions**

Yunyun Sun was a major contributor in writing the manuscript. Peng Li gave suggestions on the structure of the manuscript and participated in modifying the manuscript.All authors read and approved the final manuscript.

**Authors' information**

Yunyun Sun(School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China), Peng Li(School of Computer Science, Nanjing University of Posts and Telecommunications & Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, 210023, China), Yutong Liu(School of Computer Science, Nanjing University of Posts and Telecommunications Nanjing, 210023, China) and Zhaohui Jiang(School of Information and Computer Science, Anhui Agricultural University, Hefei, 230036, China).

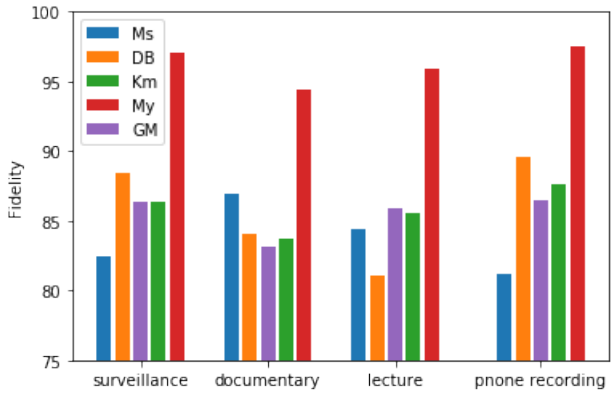


Figure 6: The fidelity results of cluster based methods on different datasets

**Author details**
[1]School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China. [2]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China. [3]Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210023, China. [4]School of Information and Computer Science, Anhui Agricultural University, Hefei 230036, China.

**References**

1. Money, A.G., Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art. Journal of Visual Communication and Image Representation **19**(2), 121–143 (2008)
2. Huang, C., Wang, H.: A novel key-frames selection framework for comprehensive video summarization. IEEE Trans. Circuits Syst. Video Technol. **30**(2), 577–589 (2020)
3. Mehmood, I., Mehmood, I., Rho, S., Baik, S.W.: Divide-and-conquer based summarization framework for extracting affective video content. Neurocomputing **174**, 393–403 (2016)
4. Song, G.H., Ji, Q.-G., Lu, Z.-M., Fang, Z.-D., Xie, Z.-H.: A novel video abstraction method based on fast clustering of the regions of interest in key frames. AEU - International Journal of Electronics and Communications **68**(8), 783–794 (2014)
5. Rachida, H., Abdessamad, E., Karim, A.: Mskvs: Adaptive mean shift-based keyframe extraction for video summarization and a new objective verification approach. Journal of Visual Communication and Image Representation **55**, 179–200 (2018)
6. Ciocca, G., Schettini, R.: Erratum to: An innovative algorithm for key frame extraction in video summarization. J. Real Time Image Process. **8**(2), 225 (2013)
7. Chang, H.S., Sull, S., Lee, S.U.: Efficient video indexing scheme for content-based retrieval. IEEE Transactions on Circuits and Systems for Video Technology **9**(8), 1269–1279 (1999)
8. Kuanar, S.K., Panda, R., Chowdhury, A.S.: Video key frame extraction through dynamic delaunay clustering with a structural constraint. Journal of Visual Communication and Image Representation **24**(7), 1212–1227 (2013)
9. Jiang, W., Fei, M., Song, Z., Mao, W.: New fusional framework combining sparse selection and clustering for key frame extraction. Iet Computer Vision **10**(4), 280–288 (2016)
10. Jeong, D.J., Yoo, H.J., Cho, N.I.: Consumer video summarization based on image quality and representativeness measure. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 572–576 (2015)
11. Liu, H., Hao, H.: Key frame extraction based on improved hierarchical clustering algorithm. In: International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (2014)
12. Wu, J., Zhong, S.-h., Jiang, J., Yang, Y.: A novel clustering method for static video summarization. Multimedia Tools and Applications **76**(7), 9625–9641 (2017)
13. Yin, Y., Thapliya, R., Zimmermann, R.: Encoded semantic tree for automatic user profiling applied to personalized video summarization. IEEE Transactions on Circuits and Systems for Video Technology **28**(1), 181–192 (2018)
14. E, M., Clemente, B, H., C, B.: Sports video summarization based on motion analysis. Computers and Electrical Engineering **39**(3), 790–796 (2013)
15. Ejaz, N., Mehmood, I., Baik, S.W.: Efficient visual attention based framework for extracting key frames from videos. Signal Processing Image Communication **28**(1), 34–44 (2013)
16. Zheng, P., Shuai, L., Kumar, S.A., Khan, M.: Visual attention feature (vaf) : A novel strategy for visual tracking based on cloud platform in intelligent surveillance systems. Journal of Parallel and Distributed Computing **120**, 182–194 (2018)
17. Zhang, Y., Liang, X., Zhang, D., Tan, M., Xing, E.P.: Unsupervised object-level video summarization with online motion auto-encoder. Pattern Recognit. Lett. **130**, 376–385 (2020)
18. Zhou, H., Sadka, A.H., Swash, M.R., Azizi, J., Sadiq, U.A.: Feature extraction and clustering for dynamic video summarisation. Neurocomputing **73**(10-12), 1718–1729 (2010)

19. Xiao, K., Lingfeng, S., Wenzhong, G., Dewang, C.: Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network. IEEE Transactions on Intelligent Transportation Systems **20**(6), 2157–2170 (2019)
20. Koch, T., Vazquez-Vilar, G.: A rigorous approach to high-resolution entropy-constrained vector quantization. IEEE Trans. Inf. Theory **64**(4), 2609–2625 (2018)
21. Liu, D., Hua, G., Chen, T.: A hierarchical visual model for video object summarization. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(12), 2178–2190 (2010)
22. Fahy, C., Yang, S., Gongora, M.: Ant colony stream clustering: A fast density clustering algorithm for dynamic data streams. IEEE Transactions on Cybernetics **49**(6), 2215–2228 (2018)
23. Niu, J., Huo, D., Wang, K., Tong, C.: Real-time generation of personalized home video summaries on mobile devices. Neurocomputing **120**(23), 404–414 (2013)
24. Yang, C., Chuang, L., Lin, Y.: Epistasis analysis using an improved fuzzy c-means-based entropy approach. IEEE Trans. Fuzzy Syst. **28**(4), 718–730 (2020)

# Figures

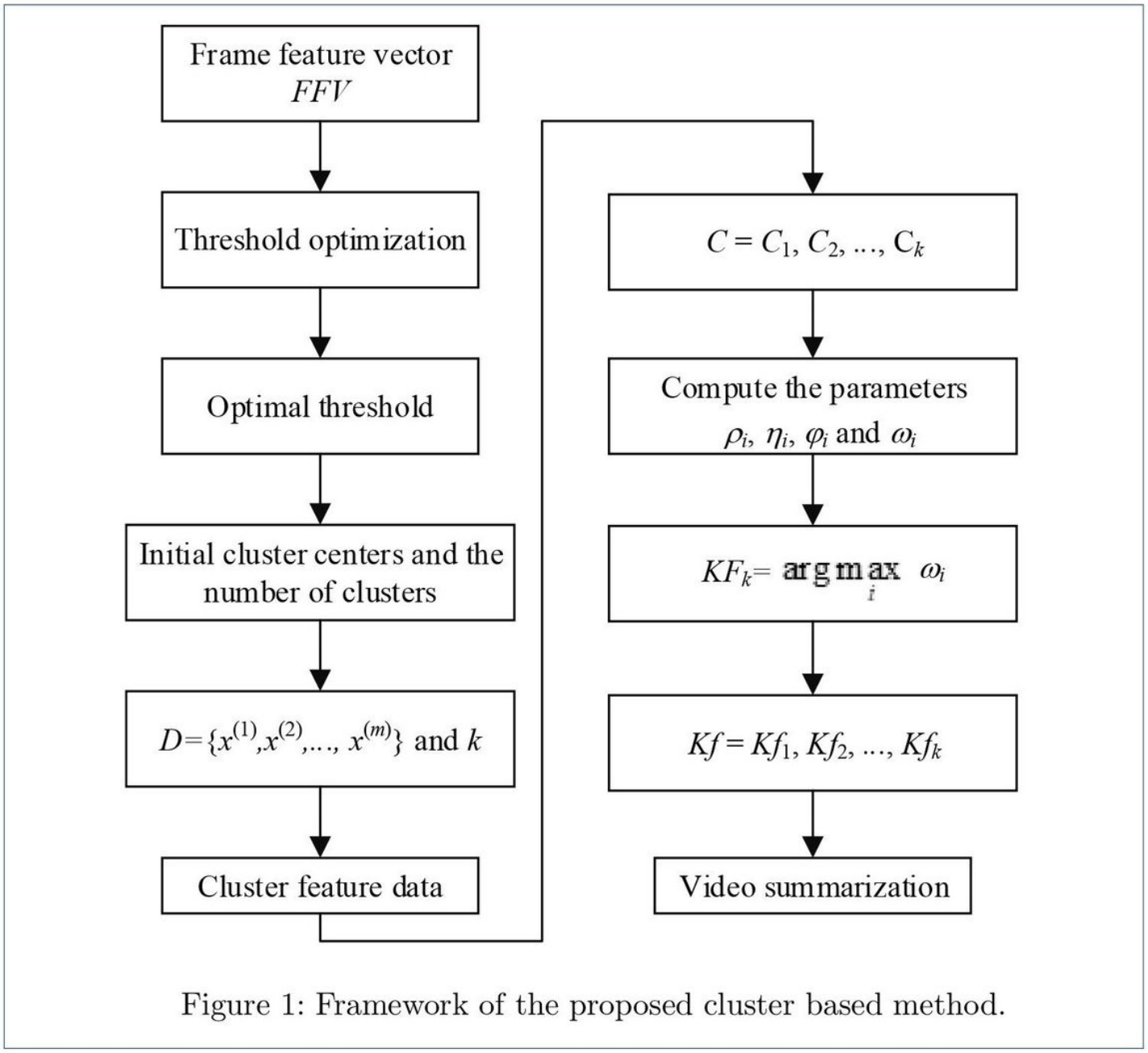


Figure 1: Framework of the proposed cluster based method.

**Figure 1**

Framework of the proposed cluster based method.

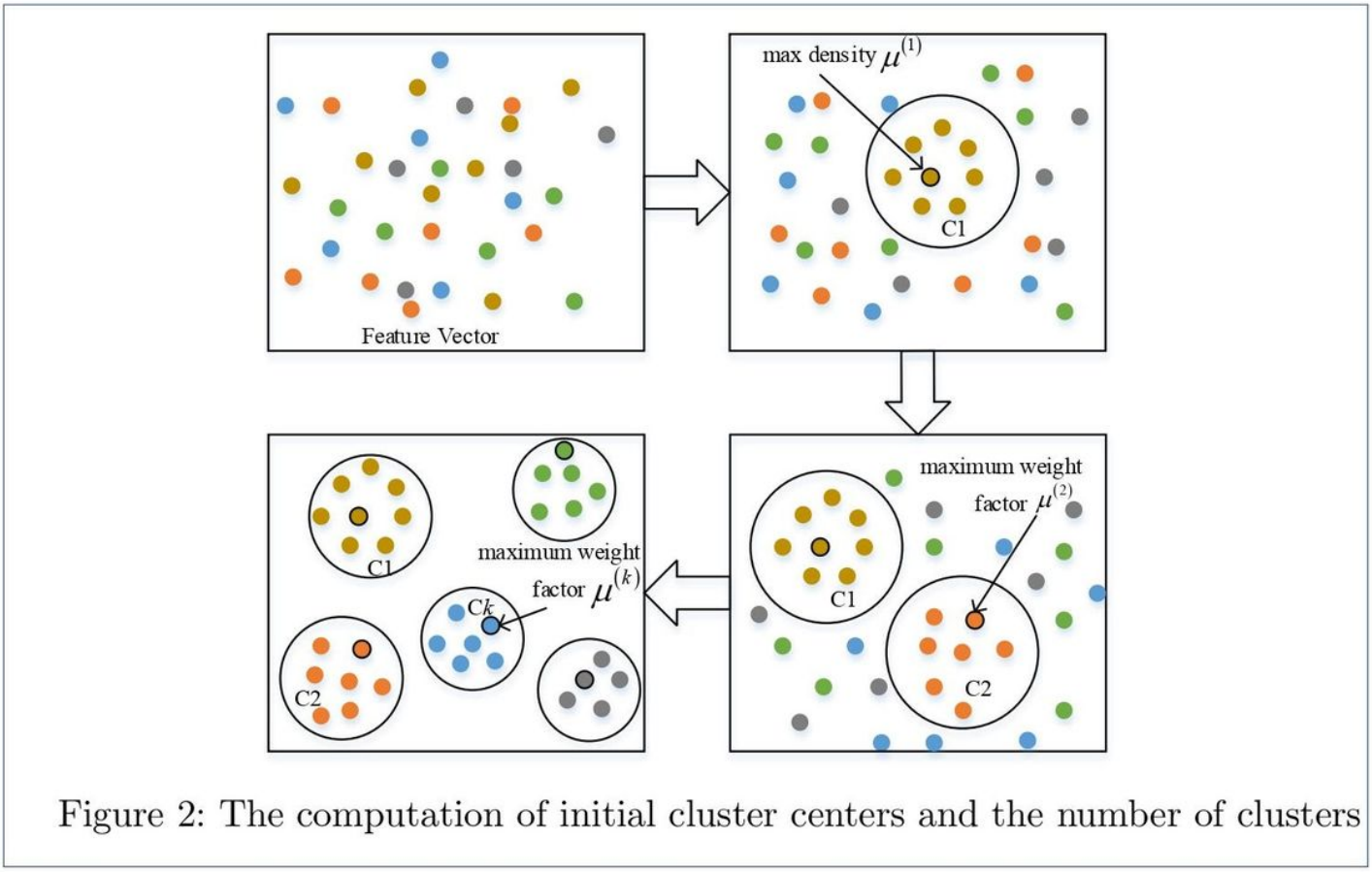


Figure 2: The computation of initial cluster centers and the number of clusters

## Figure 2

The computation of initial cluster centers and the number of clusters

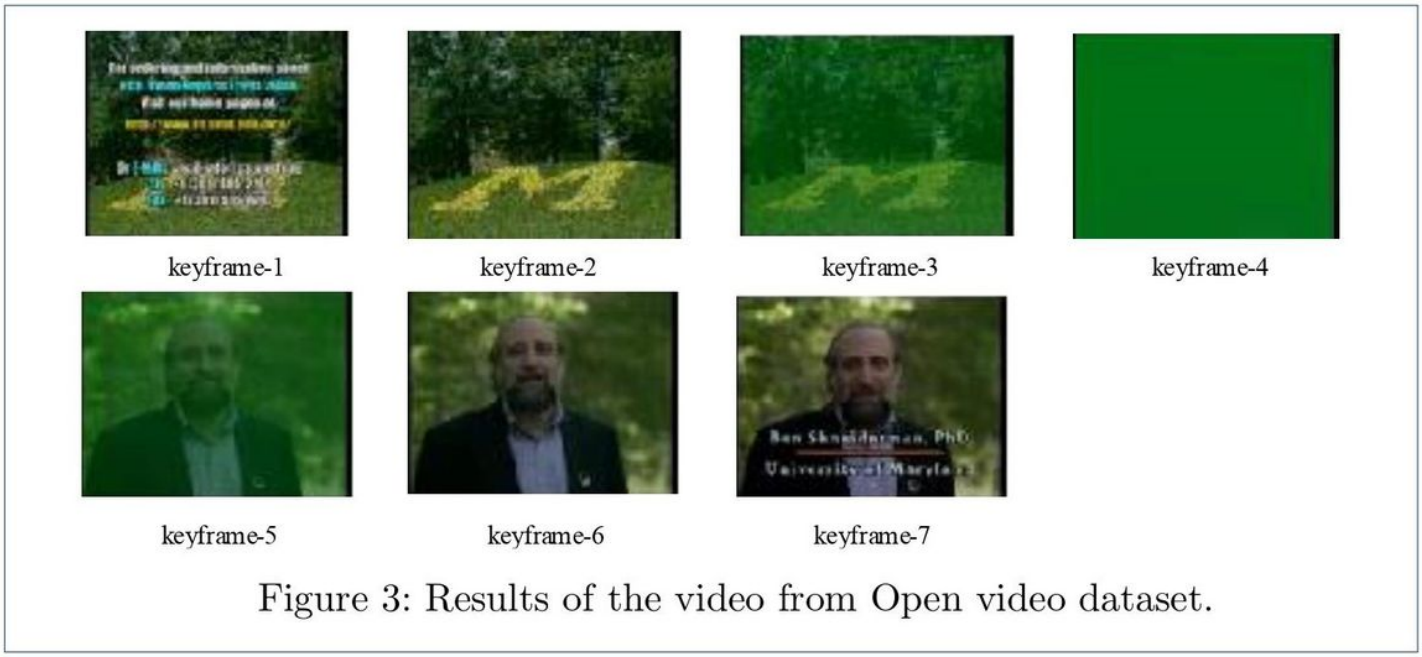


Figure 3: Results of the video from Open video dataset.

## Figure 3

Results of the video from Open video dataset

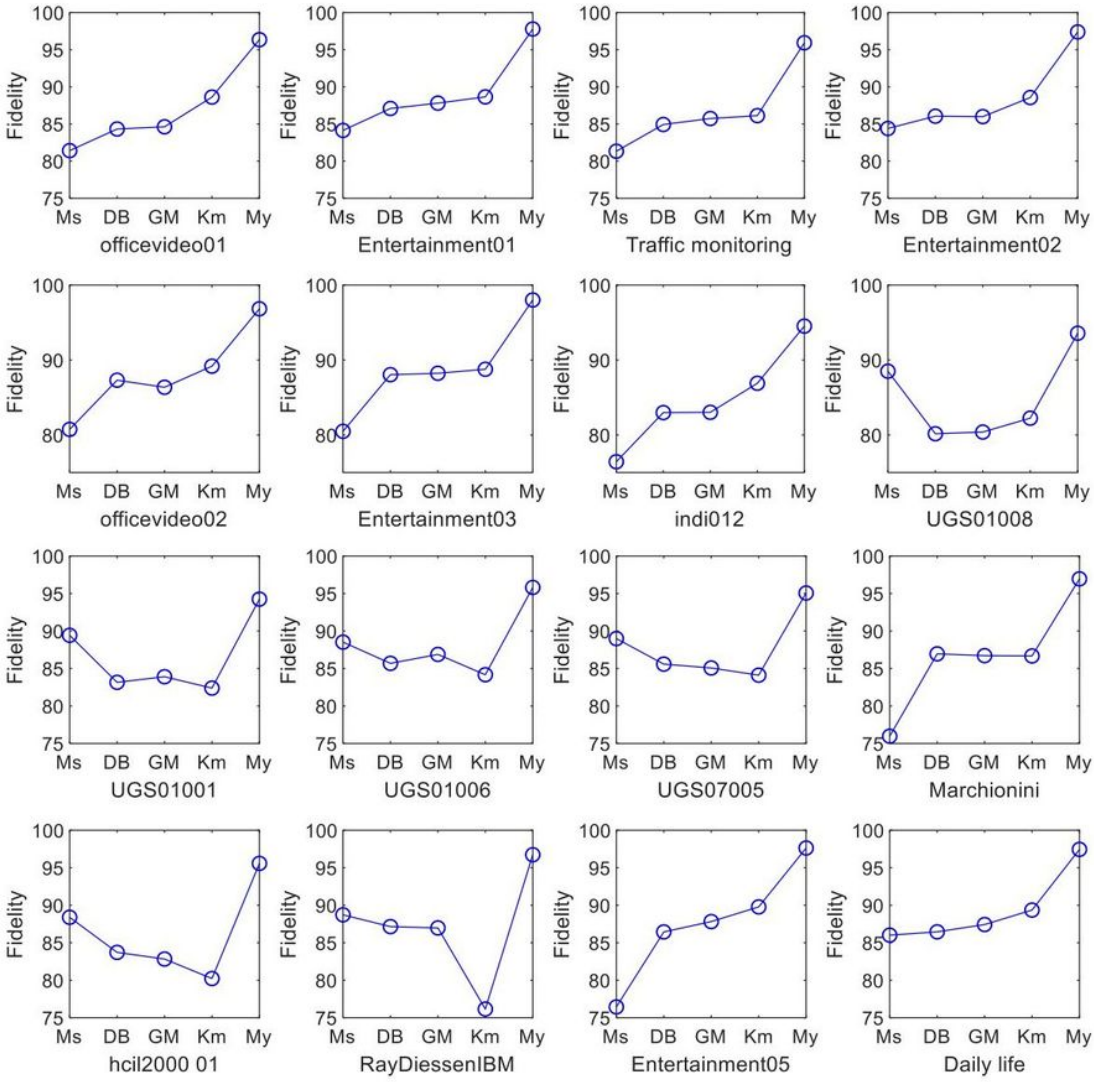


Figure 4: The fidelity measure performance of cluster based methods

## Figure 4

The fidelity measure performance of cluster based methods

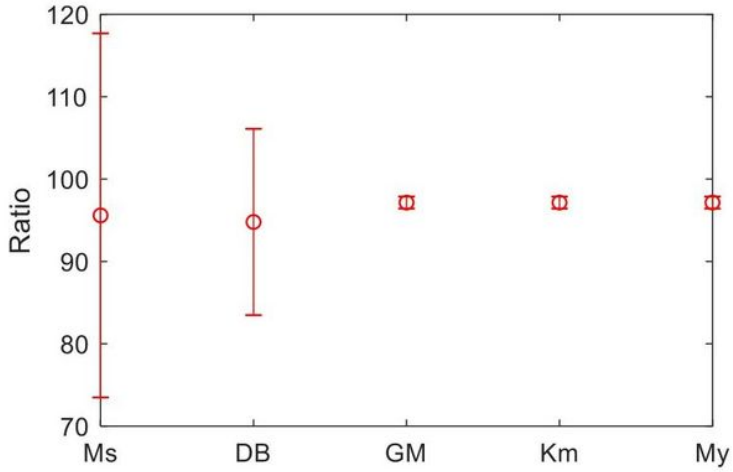


Figure 5: The ratio error-bar of cluster based methods

## Figure 5

The ratio error-bar of cluster based methods

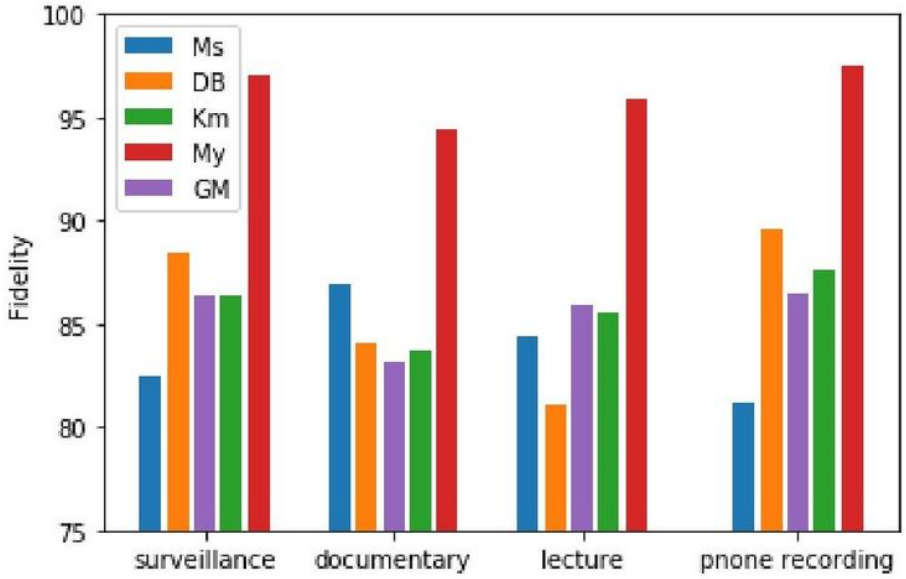


Figure 6: The fidelity results of cluster based methods on different datasets

## Figure 6

The fidelity results of cluster based methods on different datasets