Journal of Information Hiding and Multimedia Signal Processing ©2021 ISSN 2073-4212
Ubiquitous International Volume **14**, Number **3**, September **2023**

A Novel Key-Frame Extraction Approach for Semantic Video Processing

Chung-Ming Kuo[a*], Chia-Jui Hsu[a], Yuan-Xin Zheng[a], Hueisch-Jy Ding[b*]

[a]Department of Information Engineering
[b]Department of Medical Imaging and Radiological Sciences
I-Shou University
No.1, Sec. 1, Syuecheng Road, Dashu Township 840, Kaohsiung, Taiwan, R.O.C.
kuocm@isu.edu.tw

*Corresponding author: Chung-Ming Kuo, Hueisch-Jy Ding

Received November, 2022, revised February, 2023

Abstract. *Key frame is a very concise representation of a video program. The main intension and brief contents of the video can be learned from the key frames. In addition, key frames may provide important features (color, shape, texture, etc.) for semantic video processing such as indexing, retrieval and summarization. Generally speaking, a video may contain both static and dynamic segments. It may not be a reliable method to extract the representative frames from video using solely static or dynamic features. In this paper, we propose a novel approach for key frame selection. According to the global motion in a video shot, video frames are classified into two categories, static segment and dynamic segment. A static key frame selection method (ETMOF) is developed for static segments. Meanwhile, a dynamic key frame selection method is utilized for dynamic segments. Experimental results show that our approach is able to extract most representative frames form both static and dynamic segments with a concise number of frame.*

Keywords: Key frame, Semantic Video Processing, Static Segment, Dynamic Segments

1. Introduction.

Nowadays, video content (TV news, variety shows, sports, movies, educational programs...) can be obtained from almost everywhere, and watching videos has become an essential part of the majority's daily life. Managing and using such a great amount of video content has become a real challenge. Therefore, powerful management tools that can efficiently search, browse, and retrieve video content from various video sources are in great demand. There is a large amount of visual content redundancy among video frames. One way to reduce this redundancy is to select representative frames, called key frames, and store these key frames for browsing, indexing, and retrieving. By browsing these key frames, one can easily understand what happens in this video. While selecting key frames manually may achieve the best results, it is not a feasible solution for large databases. Automatically analyzing videos and selecting key frames by computers is a common goal for researchers. Thus, efficiently extracting key frames from various videos has become one of the most important research topics in semantic video processing [1-4]. Many researchers have been working on the related issue for a long time, and many techniques have been proposed. Zhang et al [5] proposed a key frame selection algorithm based on feature difference. The first frame of a shot is selected as the first key frame of the shot. Then, the difference of color histograms between the current frame and the key frame is computed. If the difference is greater than a pre-defined threshold, then the current frame is set to be a new key frame. The drawback of this algorithm is that when the change rate is small, which means the content does not change very much from frame to frame, the selected key frame will always be the first frame only. Although the change rate is small, there is no guarantee that the first frame will be very similar to the last frame in the shot. On the contrary, for most cases, the first frame is only an initiation, and lots of things will gradually appear. Therefore, the first frame selected as the only key frame for that shot seems incompletely representative of that shot. Later on, Hanjalic et al [6] proposed a key frame extraction method based on clustering. The entire video material is first grouped into N clusters, where N is the optimal combination(s) of clusters found by applying the cluster-validity analysis, each containing frames of similar visual content. Then, each of the clusters is represented by one characteristic frame, which then becomes a new key frame of a video sequence. The drawback of this method is that both the clustering and the cluster-validity analysis are time-consuming. Liu et al [7] proposed a key frame extraction method based on a triangle model of perceived motion energy (PME). With this model, a video shot is segmented into sub-segments of consecutive motion patterns in terms of acceleration and decelerations. The left-bottom vertex of the triangle represents the start point of the motion acceleration process, the right-bottom vertex represents the end point of the motion deceleration process, and the top vertex of the triangle represents the point of the maximum speed and is selected as a key frame. If there are no motion patterns in a shot, the first frames of the shot detected by the color-histogram-based algorithm are selected as the only key frames. This method is good for dynamic shots, but for those relatively static shots, it suffers from the same problem as Zhang's. Suz et al [8] proposed a key frame representation scheme, called temporally maximum occurrence frame (TMOF), based on global statistics. In TMOF, a shot is represented by a constructed frame, whose value at each pixel position corresponds to that of the pixel with the largest probability of occurrence. This provides the best results for retrieving but not for browsing, because the representative frame is a synthetic image and not real for human inspection. There are some methods of key-frame extraction that are based on clustering algorithms. The main idea is to classify each frame into various categories by using a specific clustering method and then select images as the key frame in each category [9-12]. Such approaches are straightforward and easy to implement. However, it requires knowledge of many parameters, such as clustering number or clustering radius, and thus the practical applicability is restricted. Recently, the developed approach is to use deep learning networks that can achieve state-of-the-art performance for many visual applications, including key-frame extraction. Nevertheless, deep learning models usually need huge amount of datasets for training, which is very expensive for human annotation [13-14]."

Previous researches mostly focus only on either the static type or the dynamic type of video. Static methods have good performance for static type of video, but have poor performance for dynamic type of video, whereas dynamic methods have good performance for dynamic type of video, but have poor performance for static type of video. To handle videos with mixed types, we propose a two stages solution. In the first stage, we partition a video shot into static portions and dynamic portions according to the global motion in the shot. In the second stage, we utilize the static method and the dynamic method to extract key frames from the corresponding portions, respectively. Therefore, our method extract very good representative frames for both static and dynamic type of shot, and the representative frames are good for indexing, browsing and retrieving.

This paper is organized as follows: Section 2 describes the method of global motion estimation (GME). Section 3 explains the basic idea of the proposed algorithm. Section 4 describes the experiments and demonstrates the experimental results. Section 5 concludes the paper.

2. **Global motion estimation [15].** Form photographer's point of view, the goal for changing focus to somewhere or something else is to catch other important information for audiences. Therefore, camera motion (global motion) reveals much more underlying information rather than merely a simple movement and is a good cue to segment video shots. To figure out the motion type of a frame, we compute the global motions frame by frame with the affine model. Then, the obtained parameters are used to classify the motion types. The affine model may be written as.

$$\begin{aligned} x'_t &= a + bx'_{t-1} + cx'_{t-1} \\ y'_t &= d + ey'_{t-1} + fy'_{t-1} \end{aligned} \tag{1}$$

where
(x'_{t-1}, y'_{t-1}):a pixel (x,y) at reference frame
(x'_t, y'_t):new position of pixel (x,y) at target frame
(a, d): translation parameters
(b, f): scale parameters
(c, e): rotation parameters
According to our extensive experiments, we found that the three pair of parameters, translation, scale and rotation, are with consistency of value. In other words, when the values of translation parameters are large the values of scale and rotation parameters are prone to large, and vice versa. This enables us to label a frame as static or dynamic based on the transition parameters only and without degradation of correctness. Furthermore, the complexity of computation is also reduced.

3. **Proposed key frames selection algorithm.** Figure 1 shows the diagram of our key frame selection algorithm. This algorithm focuses on key frame extraction within a given shot and the number of key frame extracted is determined by the content itself. For a given shot, the first stage is frame type determination. Then, a static and a dynamic key frame extraction method is used to extract the key frame(s) for the corresponding portions. We will detail our method on the next three subsections.

3.1. **Static and dynamic segment determination.** In section 2, we have briefly described the process of global motion estimation. Only two translation parameters are used for the determination of frame type. When the sum of the two translation parameters is greater than a predefined threshold, the frame is labelled as a dynamic frame otherwise a static frame. As shown in upper portion of figure 1, a frame type array is created to store the frame type of each frame. The frame types may fluctuate form frame to frame owing to noise. Therefore, after every frame of the entire shot has been processed, a medium filter is applied to the frame type array to smooth the distribution of the frame types. Then, the length of each consecutive segment with same frame type is counted. If the length of a segment is less than a predefined minimum length then the segment is discarded. Figure 2 depicts the approach of discard.

3.2. **Key frame extraction for static segment.** For a static segment, we developed a static method based on the TMOF to extract key frame. TMOF is an optimal algorithm to extract key frame from static segment for retrieving because the extracted key frame has the minimum sum of differences (SOD), (i.e., $SOD_{key} = \sum_{j=1}^{N} diff(f_j, f_{key})$, *the sum of the frame difference* between key frame and all other N frames in a dynamic segment) which means the most similar to all other frames in the same segment. Nevertheless, it is no good for browsing because the extracted key frame is a synthetic image and not real for human inspection. To address this problem, we proposed an extended version of TMOF, called ETMOF. First, the TMOF is computed for a static segment. Second, for

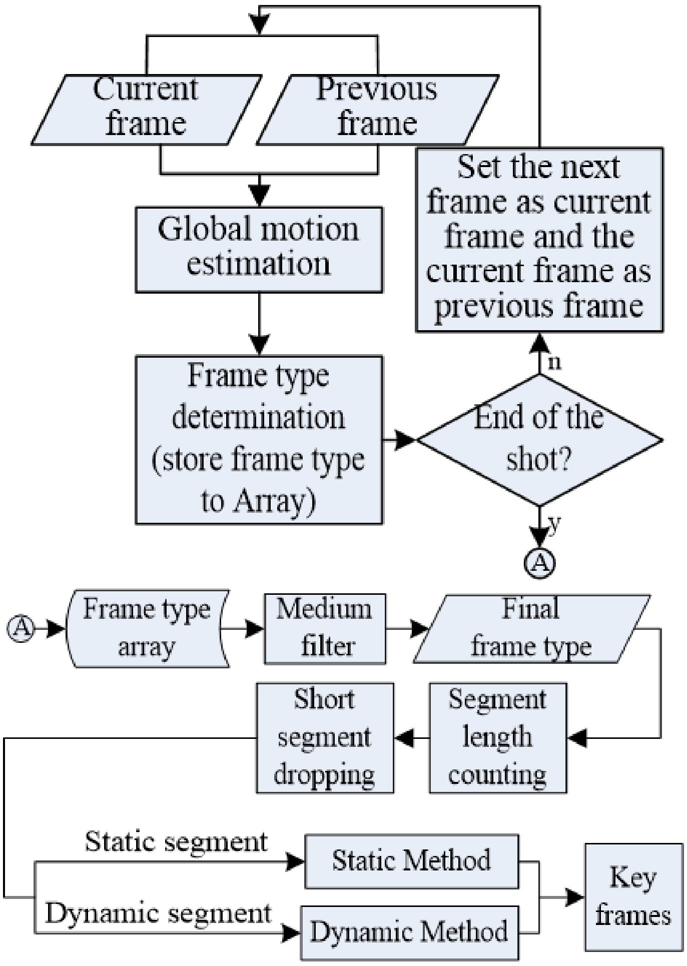


Figure 1. Diagram of proposed key frame selection algorithm.

each frame in that segment the distance between the frame and the constructed frame of TMOF is calculated. Finally, the frame with minimum distance is selected as the key frame. We may express the procedure mentioned above using a simple equation as follows.

$$key\,frame = \overset{\arg\min}{f_i \in S} \left(MSE \left(TMOF, f_i \right) \right) \tag{2}$$

where
S: the segment to be processed.
f_i: a frame in S.
$MSE(.)$: a function, which computes the mean square error between two image.

3.3. **Key frame extraction for dynamic segment.** The PME model achieves good results for dynamic segments. Therefore, we adopt PME to extract key frames from dynamic segment [7]. A dynamic segment usually consists of motion with a gradual acceleration to a peak, followed by a gradual deceleration. Fig. 3 shows the motion energy curve of each frame in a video shot. There are three triangles in this figure, which means that there are three motion events in this shot, because one motion event will be represented by a triangular model.

In our work, we use motion energy to extract key frames. By calculating the motion energy of each frame, we can obtain a motion energy curve over time. From the curve, the entire segment over time can be evaluated by the change of motion intensity. In our

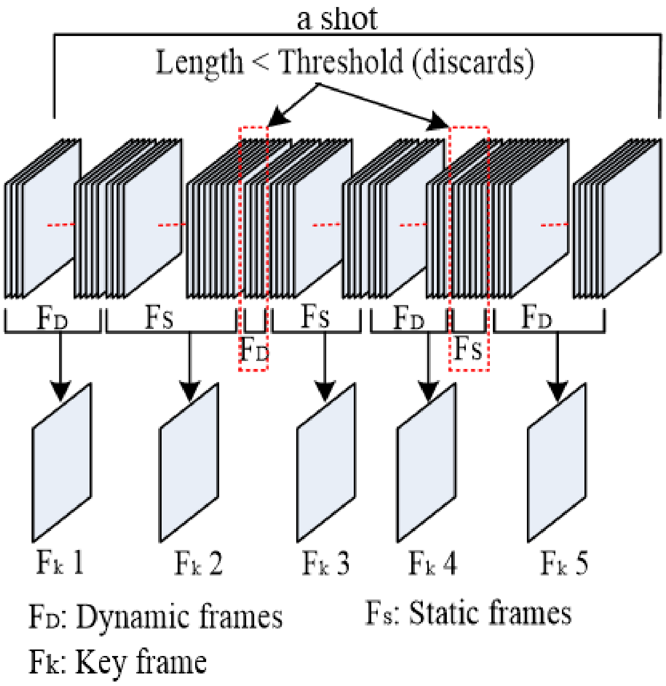


Figure 2. Diagram of short segment dropping and key frames extraction.

work, for dynamic segments, we use motion estimation to calculate the motion energy. In the energy curve, the frame location with the maximum motion energy is selected as a key frame of the dynamic segment.

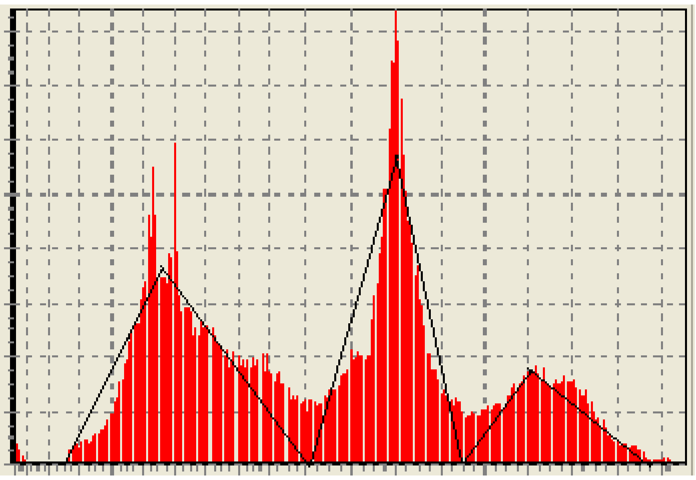


Figure 3. Motion energy triangle model

Next, we explain the calculation method of the motion energy. The flow chart is shown in Fig.4. The video frame is first divided into $B \times B$ blocks, i.e., the same as traditional motion estimation, the total blocks of a frame can be represented as N. The $MixFEn_{i,j}(t)$ is the motion intensity that calculated by motion vector with previous frame, and $MixBEn_{i,j}(t)$ is calculated by next frame, respectively. We define $\alpha(t)$ as the direction of most significant motion intensity. The directions of motion vector are distributed in 2π for simplicity we divided into sixteen directions. And then we count the histogram of motion directions of each block motion vector in sixteen directions. Therefore

$\alpha(t)$ can be calculated as following

$$\alpha(t) = \frac{\max\left(AH\left(t,k\right), k \in [1,n]\right)}{\sum_{k=1}^{n} AH\left(t,k\right)}, \tag{3}$$

where $AH(t,k)$ is the k^{th} bin of n directional histogram, i.e., in our work n equals sixteen. And motion intensity is represented as

$$Mag\left(t\right) = \frac{\left(\frac{\sum \mathrm{Mix}FEn_{i,j}(t)}{N} + \frac{\sum \mathrm{MixB}En_{i,j}(t)}{N}\right)}{2}, \tag{4}$$

$$PME(t) = Mag(t) \times \alpha(t) \tag{5}$$

where N is the total number of blocks in each frame. Finally, as shown in Fig, 5, the key frame is extracted from the frame with the most significant motion energy.

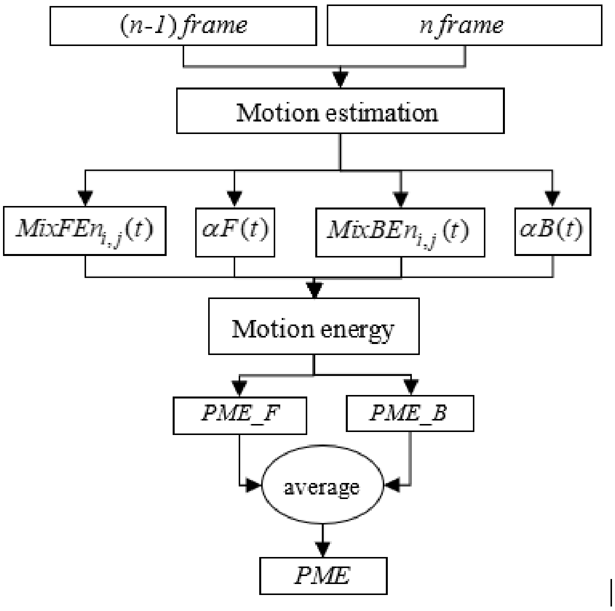


Figure 4. The flow chart of calculation of motion energy

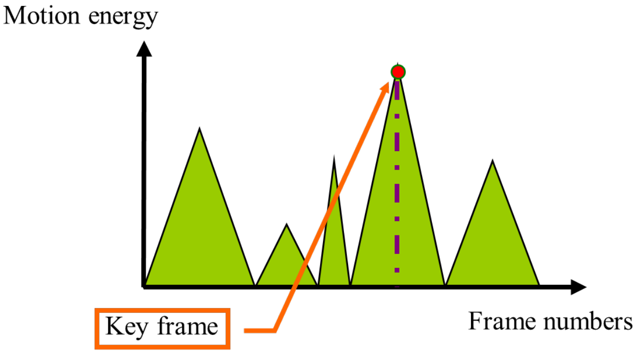
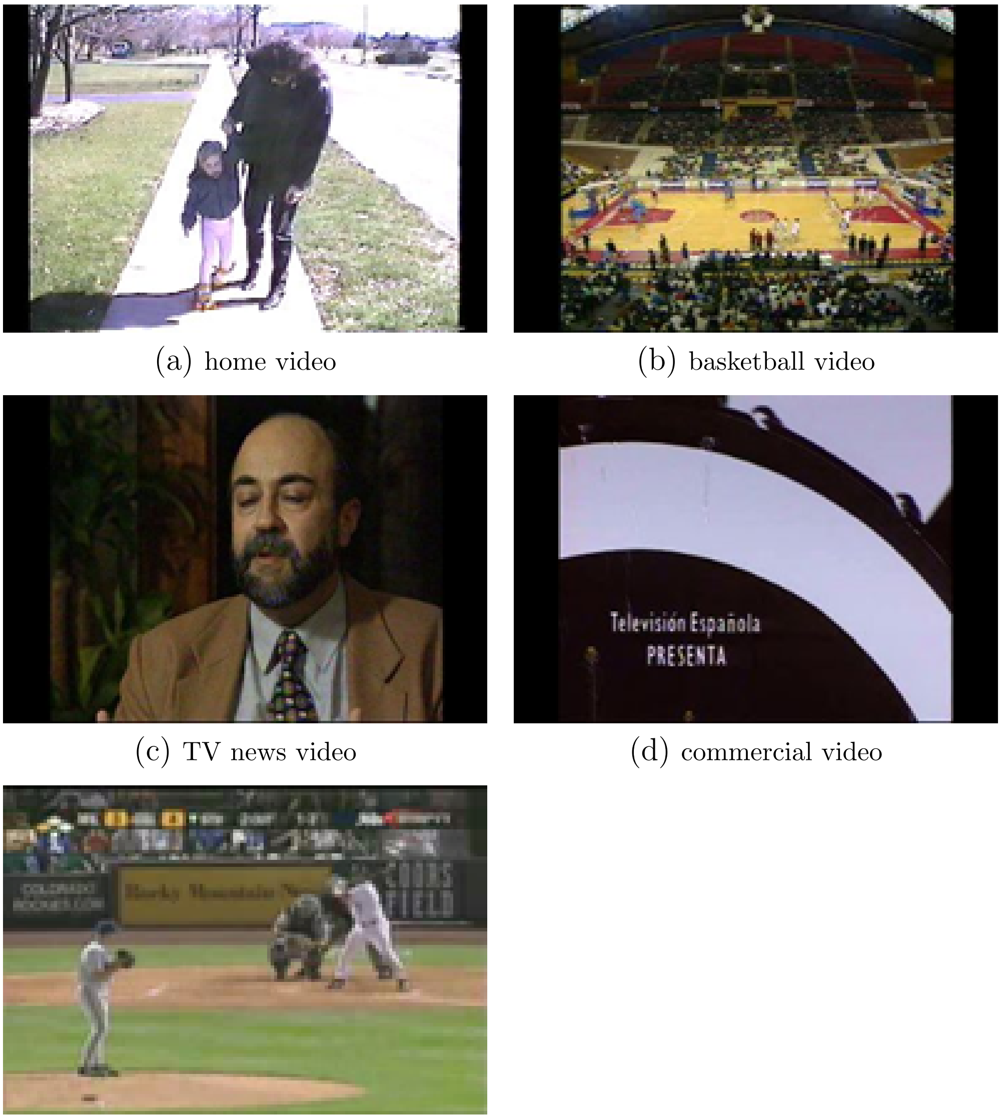


Figure 5. Motion energy curve of frames

4. **Experimental results.** Because of the subjectivity of key frame extraction issue, no ground true database exists for testing. Hence, we select four video from MPEG-7 database (including a home video, a basketball video, a TV news video, and a commercial video) and one baseball video recorded by ourselves as the test samples, shown in Table 1. To evaluate the performance of our algorithm, ten subjects are invited to join our evaluation.

Table 1. Test videos.



(a) home video

(b) basketball video

(c) TV news video

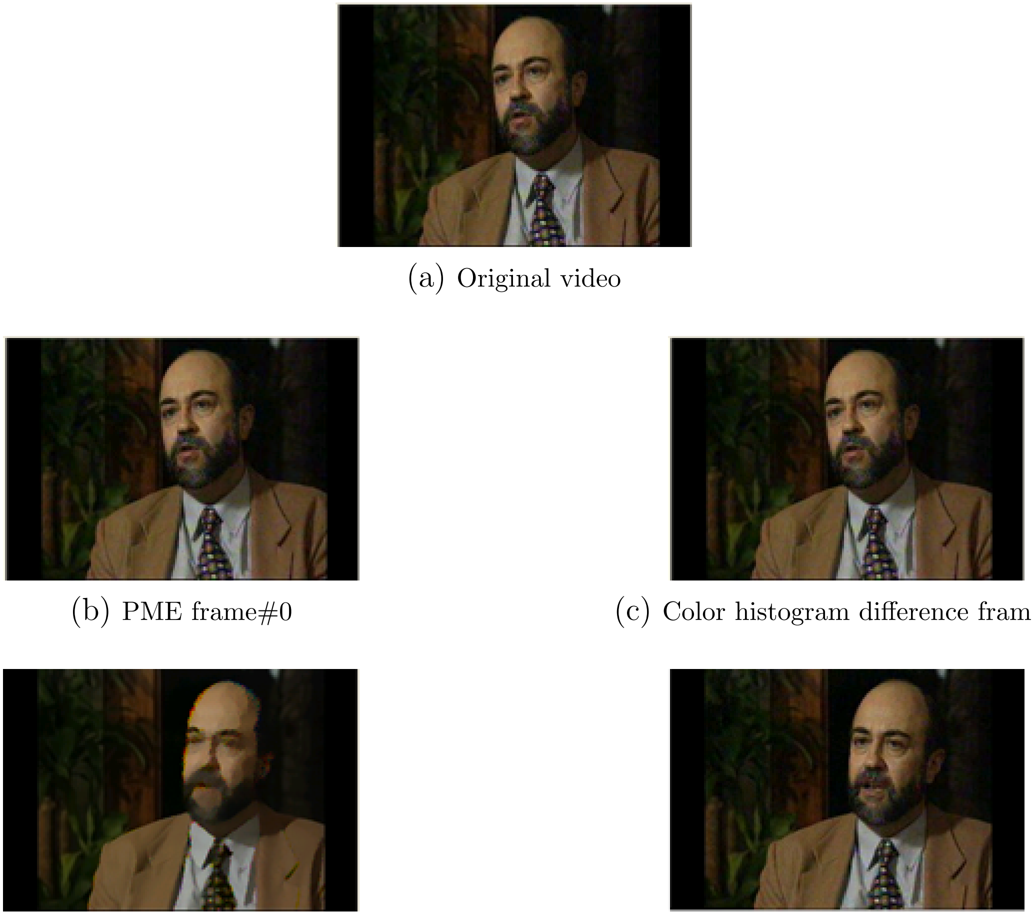
(d) commercial video

(e) baseball video

As described in Section 2, we use the variation of global motion parameters and predefined minimum length to divide the video into static segment and dynamic segment, and then the key frame is extracted respectively by proposed static method and PME method according to video types. In our work, we set the minimum length is 15, and the translation of global motion is larger than the average block motion. In the following, we compare the performance with some state-of-the-art methods, i.e., TMOF, PME and color histogram difference, to demonstrate the superiority of proposed method.

For static segment, two test video shots are selected for comparison. A shot of "TV news video" and "home video" are with 368 frames and 335 frames, respectively. The testing shot is the sequence with very small global motion, which is suitable for us to use for the experiment of capturing key frames for the static segment. The comparisons of various static key frame extraction methods are shown in Fig. 6 and 7. From the experimental results, we can find that the traditional method of color histogram difference always select the first frame of the shot as the key frame. The TMOF algorithm is only suitable for almost static shots, and the key frames selected from moving shot are not satisfactory. However, for proposed method, the extracted key frame not only with small SOD but also with representativeness. Therefore, we believe that our static method is superior to that of other methods.

Figure 8 is the comparison of key frame extraction for dynamic segments. The testing shot consists of a Taiwan professional baseball shot with 225 frames. From the content, we can find that this is a shot with obvious global motion. Here we compare the temporal most frequent picture method (TMOF), the motion energy analysis method (PME) and our method.



(a) Original video

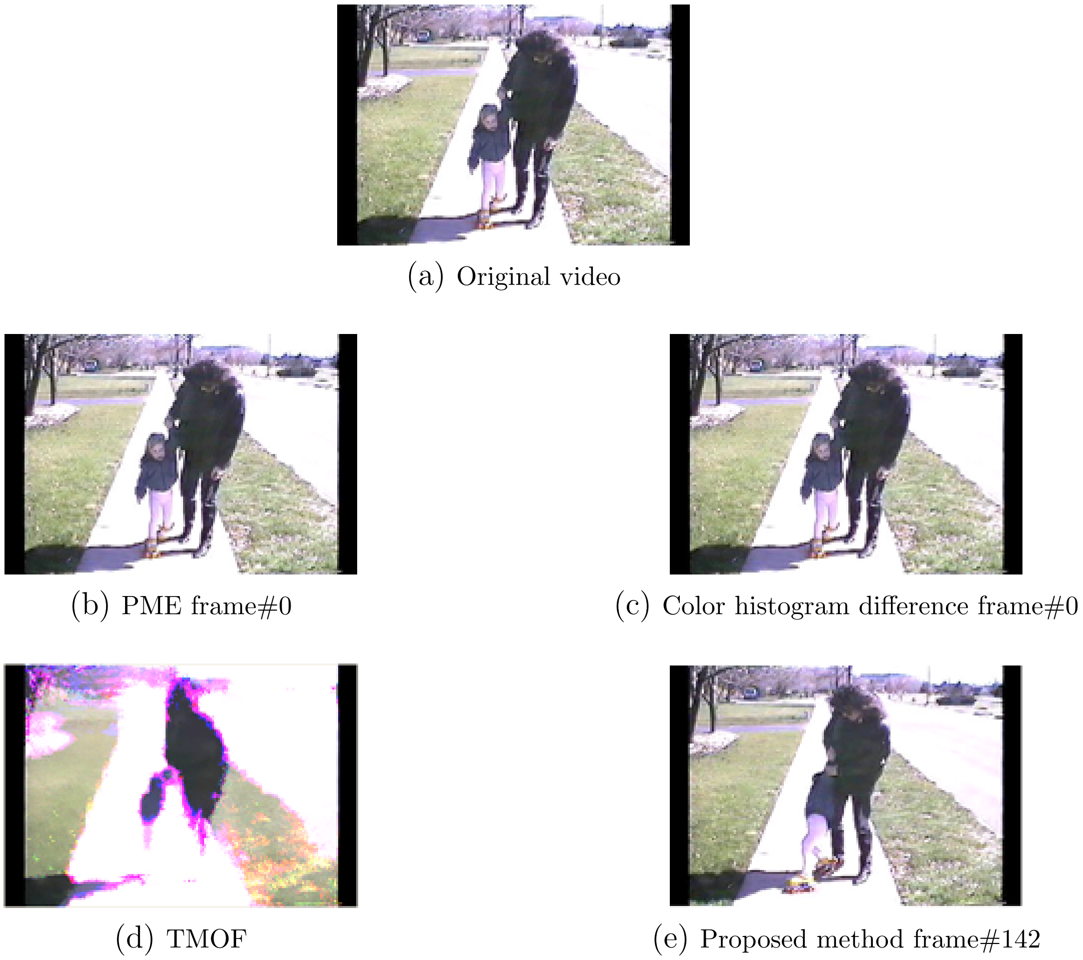
(b) PME frame#0

(c) Color histogram difference frame#0

(d) TMOF

(e) Proposed method frame#142

Figure 6. The extracted key frames from TV news video by various methods



(a) Original video

(b) PME frame#0

(c) Color histogram difference frame#0

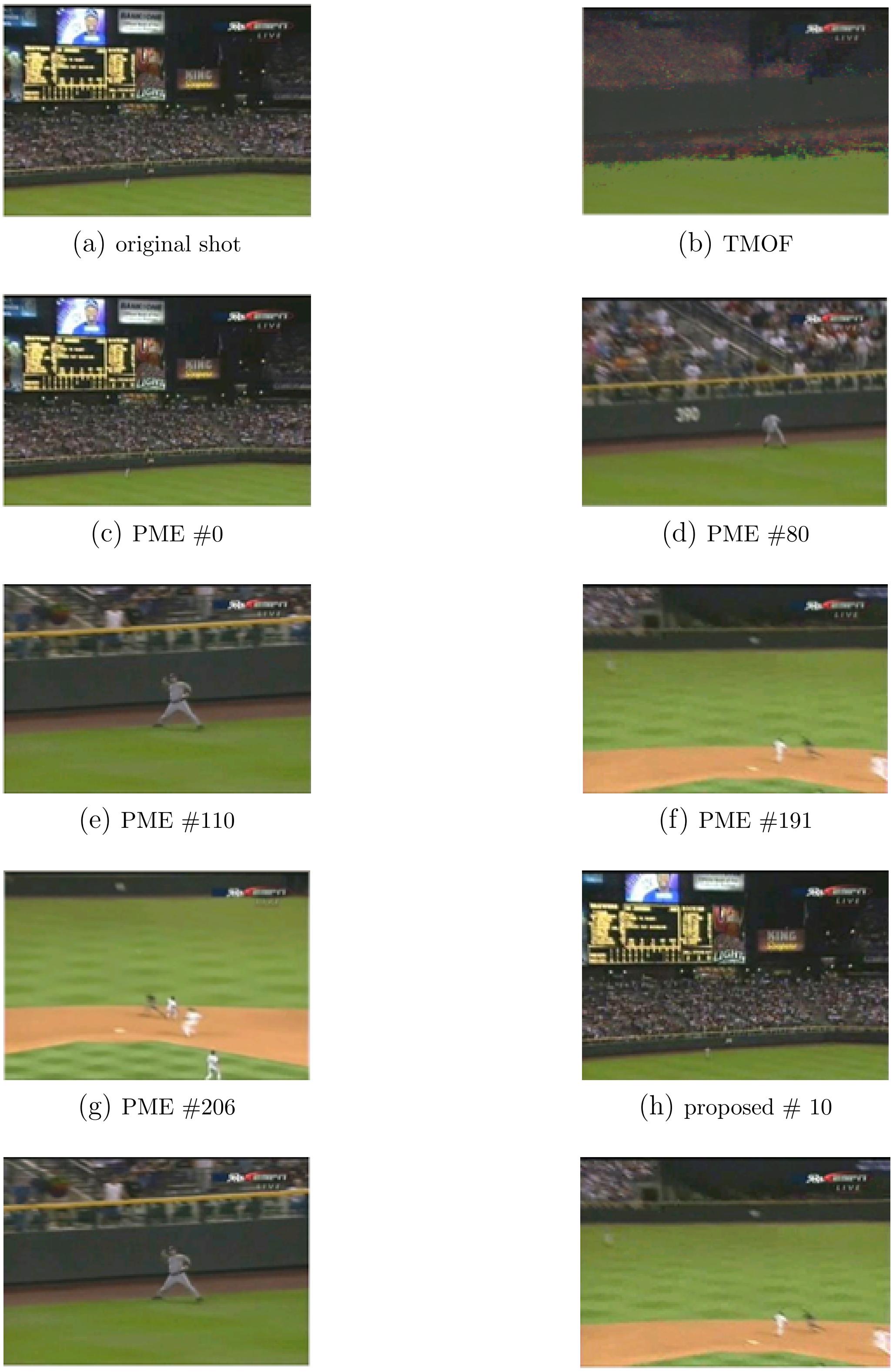
(d) TMOF

(e) Proposed method frame#142

Figure 7. The extracted key frames from home video by various methods

For TMOF, since the global motion in video shot is significant. Thus, the extracted key frame is very blurred except the subtitles in the shot. Hence, this method is not suitable for extraction key frame of dynamic segments. For PME, it usually selects more key frames, i.e., 5 key frames, than the proposed method, i.e., 3 key frames. For a long shot, it implies higher storage space and higher computational complexity. Obviously, the performance of proposed method is superior than that of conventional methods.

Finally, subjects' evaluation is performed. One hundred shots are trimmed from each test video and totally 500 shots are used for testing. The key frames extracted by PME and our method are shown with the shot in a random order. Subjects have no information about which key frames are extracted by which method for the sake of justness. They may rate the selected key frames as one of the following three levels, Good, Fair, or Poor. Table 2 and 3 show the results of the evaluation. Compare with the PME, our method has a better rating for all kind of videos. In addition, our method has a lower ratio of average key frames per shot, which implies we may use fewer key frames to represent a shot and a lower storage space is required and lower complexity for further processing.



(a) original shot

(b) TMOF

(c) PME #0

(d) PME #80

(e) PME #110

(f) PME #191

(g) PME #206

(h) proposed # 10

(i) proposed # 110

(j) proposed # 191

Figure 8. Dynamic key frames extracted from baseball video by various methods (a) shot schematic, (b) by TMOF, (c) (g) by PME, (h) (j) by proposed method.

Table 2. Average rating of the proposed method.

	Good	Fair	Poor	Shots	Average rating
Home video	92.5	5.3	2.2	100	5.6
Basketball video	84.3	14.6	1.1	100	2.1
TV news video	94.7	5.3	0	100	1.7
Commercial video	91.8	6.1	2.1	100	1.1
Baseball video	92.2	5.4	2.4	100	2.2

Table 3. Average rating of the PME.

	Good	Fair	Poor	Shots	Average rating
Home video	82.3	14.2	3.5	100	8.8
Basketballvideo	81.6	17.2	1.2	100	6.1
TV news video	77.1	21.4	1.5	100	5.8
Commercial video	71.2	25.6	3.2	100	3.1
Baseballvideo	83.7	14.3	2.0	100	3.1

5. **Conclusion.** We have proposed a novel key frame extraction algorithm. The major contribution of this paper is that we classify video shots into static and dynamic segments. This allows a proper key frame extraction method (static or dynamic) to be used to deal with the corresponding segments and therefore improve the performance of the entire system.

REFERENCES

[1] X. Liu, M. Song, L. Zhang, S. Wang, J. Bu, C. Chen, and D. Tao, "Joint shot boundary detection and key frame extraction." 21st International Conference on Pattern Recognition (ICPR2012), pp. 2565–2568, Nov. 2012.

[2] H. Liu, L. Pan; W. Meng, "Key frame extraction from online video based on improved frame difference optimization," IEEE 14th International Conference on Communication Technology, pp. 940–944, Oct. 2012.

[3] D. Tian, "Gaussian Mixture Model and Its Applications in Semantic Image Analysis," *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 9, No. 3, pp. 703–715, May 2018.

[4] J. Wang, W. Song, X. Sun, L. Tang, J.H. Yeh, "Annotation Method to Improve the Mapping Between Image Features and High Level Semantic Expression," *Journal of Network Intelligence*, Vol. 5, No. 4, pp. 211–217, November 2020.

[5] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An Integrated System for Content-based Video Retrieval and Browsing," *Pattern Recognition*, Vol. 30, No. 4, pp. 643–658, 1997.

[6] A. Hanjalic and H. J. Zhang, "An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster Validity Analysis," *IEEE Transactions on Circuits and System for Video Technology*, Vol. 9, pp. 1280–1289, Dec. 1999.

[7] T. Liu, H.J. Zhang, and F. Qi, "A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 13, No. 10, Oct. 2003.

[8] K.W. Sze, K.M. Lam, and G. Qiu, "A New Key Frame Representation for Video Segment Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 9, Sept. 2005.

[9] J.X. Wu, S.H. Zhong, J.M. Jiang, Y.Y. Yang, "A novel clustering method for static video summarization," *Multimed. Tools Appl.*, Vol. 76, pp. 9625–9641, 2017.

[10] H. Gharbi, S. Bahroun, M. Massaoudi, E. Zagrouba, "Key frames extraction using graph modularity clustering for efficient video summarization," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1502–1506, March 2017.

[11] H. Tang, H. Liu, W. Xiao, N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *Neurocomputing*, Vol. 331, pp. 424–433, 2019.

[12] Y. Chen, T. Huang, Y. Niu, X. Ke, Y. Lin, "Pose-Guided Spatial Alignment and Key Frame Selection for One-Shot Video-Based Person Re-Identification," *IEEE Access 2019*, Vol. 7, pp. 78991–79004, 2019.

[13] M. Zhao, X. Guo, X. Zhang, "Key Frame Extraction of Assembly Process Based on Deep Learning," 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 611–616, July 2018.

[14] M. Guangyi, S. Xiaoyan, "Interested Keyframe Extraction of Commodity Video Based on Adaptive Clustering Annotation," *MDPI in Applied Sciences*, Vol. 12, pp. 1–18, Jan. 2022.

[15] Y. R. Huang, C. M. Kuo, C. L. Kuo, "An Efficient Global Motion Estimation Algorithm Using Recursive Least Square," *SPIE-Optical Engineering*, vol. 45, no. 5, pp. 057003-1–057003-13, May 2006.