

# An Improved Histogram Based Image Sequence Retrieval Method

Xiang Fu, and Jiexian Zeng  
Nanchang HangKong University, Nanchang, China  
Email: fxfb163@163.com

**Abstract**—In this paper, histogram difference between local images (HDLI) is used to image sequence retrieval. Firstly, both the query image sequences and every database sequences are segmented into several shots based on HDLI of consecutive frames. Then, for each shot, one or more key frames are selected based on HDLI of benchmark frame and each followed frame. Third, to retrieve videos similar to the given query video, similarity between the key frames of query shot and each database shot is computed. The similarity is also measured using HDLI of key frames. Experimental results show that the proposed method improves the image sequence retrieval accuracy efficiently.

**Index Terms**—image sequence retrieval, shot boundary detection, key frame extraction, local image, interest points

## I. INTRODUCTION

Retrieval of digital video is a very active research area. Previous work on video retrieval can be classified into two main streams: keywords-based methods and key frames-based methods. For automatic video retrieval, it is almost impossible to use keywords to describe video sequences [1]. A generic approach is first to segment a video into shots. Then, one or more key frames can be extracted for each video shot. The key frames are then used to represent the video shots for retrieval.

After comparing most general video shot retrieval algorithms, we have found that there are 3 main components that affect the performance of retrieval: (1) Video shot boundary detection is the first step of video retrieval. (2) The key frames should provide the most representative power in video shot representation. (3) The feature matching is useful for finding contents similar to a query video stream from video database.

In this paper, we research from these three aspects. Firstly, shot boundary is detected based on HDLI of consecutive frames. It can overcome the deficiency of traditional histogram-based method that different contents frames have similar histograms, and can detect gradual changes effectively. Secondly, key frames extraction method is based on HDLI of the benchmark frame and each followed frame. Finally, shot retrieval is implemented based on HDLI of key frames for query shot and shots of video database.

## II. SHOT BOUNDARY DETECTION

There are broadly two approaches, the pixel domain processing and the compressed domain processing. Pixel domain methods include pixel-to-pixel frame comparison, block-to-block frame comparison and histogram-based frame comparison [2]. Compressed domain approaches make direct use of the encoded data. Lee et al exploit information from the first few AC coefficients in the compressed domain, and track binary edge maps to segment the video [3]. Macroblock based methods also work on compressed MPEG video [4].

Comparison of various methods has shown that the pixel domain methods have higher accuracy compared to the compressed domain methods [5,6]. The compressed domain methods, on the other hand, work faster. Also, of all the pixel domain methods, histogram based techniques perform better than the rest [2]. However, for the histogram-based methods, it is possible that the histograms of two frames are similar, but the contents are completely different [7]. In order to handle this situation, we propose a shot detection method using local features.

### A. Interest Point Detection

The Harris interest point detector[8] is widely used for its good performance and invariance to rotation and translation. Fig.1 shows examples of interest points detected by Harris. Comparing Fig.1(a) with Fig.1(b), the background has changed for the camera moving with objects, while the interest points can keep stable.

### B. Local Image Extraction

Local image is consisting of regions around interest points. As shown in Fig.2, where the value of  $R$  centered on the interest point is 30 pixels. A smaller value of  $R$  means a larger stable area will be ignored and the more of the local region be emphasized, thus the more sensitive to the difference between frames.

### C. Histogram of Local Image

For shot boundary detecting, we compute the histogram of local image obtained in last step. As shown in Fig.3, Fig.3(a) and Fig.3(b) are artificial images, the values on each block is its gray value, the largest block in the middle of Fig.3(a) has value of 110 is interpolated into other blocks in Fig.3(b). Undoubtedly, they have same histograms as shown in Fig.3(e).

---

This work was supported partially by the National Natural Science Foundation of China (Grant No. 60675022).

Corresponding author: Xiang Fu

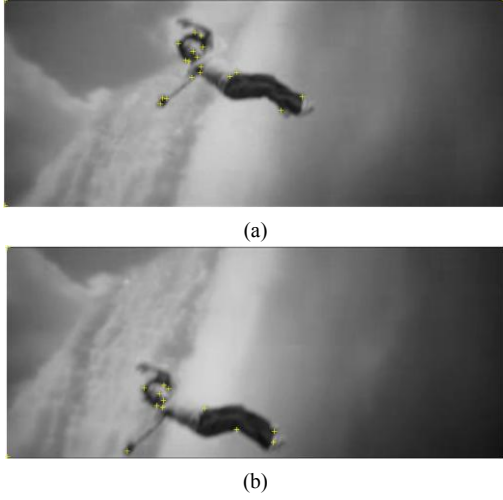


Figure 1. Examples of interest points detected by Harris detector.

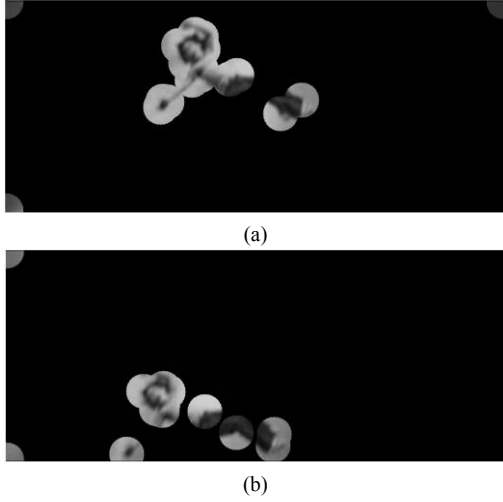


Figure 2. Local images defined by interest points in Fig. 1.

In Fig.3(a) and Fig.3(b), the interest points are described as little cross, corresponding local images are Fig.3(c) and Fig.3(d). And the histograms of two local images are shown in Fig.3(f), it can be seen that the histograms become different completely. Especially, the bin of gray value 110 for Fig.3(d) is larger than that of Fig.3(c), where a larger smooth area was ignored.

#### D. Shot Boundary Detection

It is found that for a shot cut, the change between consecutive frames is large. However, for gradual transitions, the change between consecutive frames is not large enough, but the number of change between the frames before and after the transition is large.

The HDLI of frame  $F_i$  and frame  $F_{i-1}$  is:

$$d(H_i, H_{i-1}) = \sum_{k=0}^{N-1} \|H_i(k) - H_{i-1}(k)\|^q \quad (1)$$

where  $N$  is the number of histogram bins, and  $q$  the order of norm distance which is set to 1 for its good performance and easy computation.

When perform shot detection, we do not use the histogram difference directly, but the difference of histogram difference, which is defined as:

$$Dd = d(H_i, H_{i-1}) - d(H_{i-1}, H_{i-2}) \quad (2)$$

$Dd > 0$  means the gray of frames changed more and more great, and  $Dd < 0$  means more and more slow. Shot boundary can be detected as follows:

$$\begin{cases} |Dd| > T_1 & \text{cut} \\ |Dd| < T_1 \text{ and } N_{Dd>0} > T_2 & \text{out} \\ |Dd| < T_1 \text{ and } N_{Dd<0} > T_2 & \text{in} \end{cases} \quad (3)$$

where  $T_1$  is the cut threshold,  $N_{Dd>0}$  is the number of frames that  $Dd > 0$ , and  $T_2$  is the gradual transition threshold.  $|Dd| < T_1$  and  $N_{Dd>0} > T_2$  means the current shot is going *out* slowly, and  $|Dd| < T_1$  and  $N_{Dd<0} > T_2$  means coming *in* slowly, as shown in Fig.4.

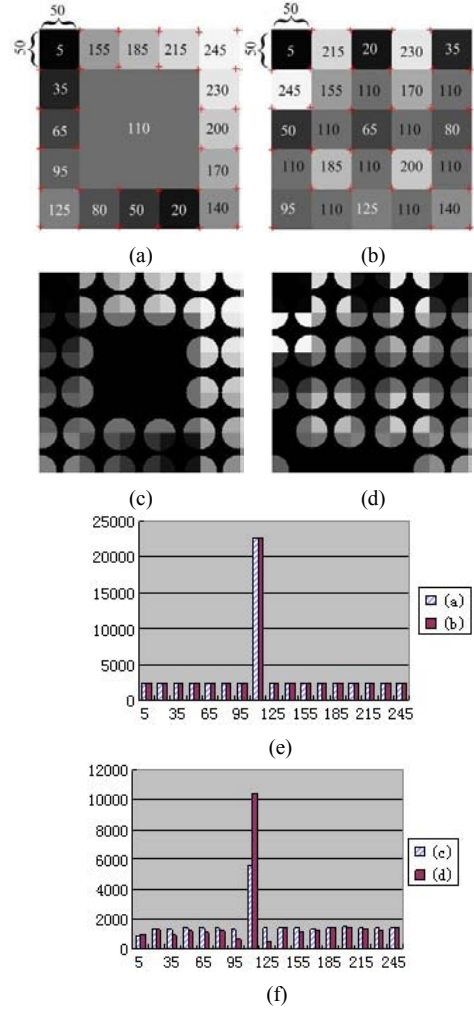


Figure 3. Histogram of original images and local images. (a) and (b) have the same gray distribution, and have the same histograms as shown in (e). Local images of (a) and (b) are shown in (c) and (d) respectively, they have different histograms as described in (f).

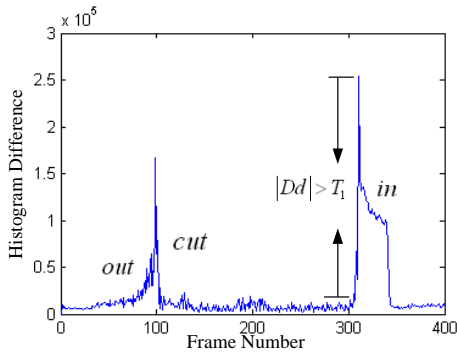
### III. KEY FRAME EXTRACTION

Key frame extraction is the process of selecting frames from shots to represent the video content. Some simple methods pick one or more key frames from every shot at predetermined temporal locations such as the first, middle and/or the last frame [9,10]. In [11] frames of a video

sequence are chosen at regular time intervals, leading to a storyboard presentation. These methods may not provide the most representative power in video shot representation, especially for shots of long duration and high motion activity. To overcome these drawbacks, in this section, an interest points based key frames extraction method is presented.

Firstly, the first frame of a video shot is selected as key frame and benchmark frame, and the followed frames compared with the benchmark frame one by one, for the difference between frame 1 and frame  $k$  is large enough, frame  $k$  is selected as new key frame. Then the followed frames compared with the new benchmark frame  $k$  one by one. Repeat the process until all the frames of a video shot processed, and different numbers of key frames are extracted for video shots of a sequence.

The key frames extraction method has good representative power in video shot representation. It is processed based on interest points, which can reduce the amount of data greatly, what's more, the histograms of local images for all frames have been obtained at the stage of shot boundary detection in Section II and can be used directly here, so the method is computationally effective.



(Dd: Difference of Histogram difference)

Figure 4. Difference of histogram difference based shot detection

#### IV. VIDEO RETRIEVAL

There are many methods for features matching have been processed in literatures. Video retrieval base on key frames is similar to static image retrieval, so the traditional technologies of image retrieval can be used. Here we also use HDLI as features for video retrieval.

As shown in Fig.5, all the key frames of query video shot will be compared with all the key frames of database video shots if there is no satisfactory result obtained.

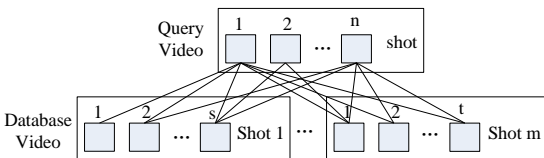


Figure 5. Process of key frames matching for video indexing

#### V. EXPERIMENTAL RESULTS

##### A. Video shot boundary detection

To evaluate the performance of the proposed shot detection method, we use various kinds of videos from

news reports and famous movies, parts of them as shown in Table I. The sequence of Earthquake Report has lots of gradual changes; Superman and Tomb Raider have many fighting scenes and the camera moving greatly; and the Friends has a lot of abrupt cuts accompanied with captions fade in and fade out.

The comparison between the proposed algorithm and traditional Histogram method relies on the recall precision and false alarms rate

$$\text{recall} = \text{detects}/(\text{detects} + \text{MD}) \quad (4)$$

$$\text{false alarms} = \text{FA}/(\text{detects} + \text{MD}) \quad (5)$$

where detects denotes the correctly detected boundaries, while MD and FA denote missed detections and false alarms. Results of recall for abrupt cuts and gradual changes detection are provided in Tables II and III, respectively. It can be seen that the performance of traditional Histogram method is close to the proposed algorithm for abrupt cuts detection, while the proposed algorithm is outperformed to the former for gradual changes. That's because the gradual changed frames have similar histograms, but have different interest points. So the local images defined based on the interest points are more sensitive to gradual changes between frames.

The large movement of camera will lead to false detection for Histogram-based method, while the proposed method is suitable for this situation, because when the camera moving with objects, most of interest points keep homologies in the regions of objects. So the false alarms rates for Histogram are much higher than that of the proposed method, as shown in Table IV.

##### B. Video retrieval

There are no objective standards to measure the performance of key frames extraction method. In this section, experimental results will be given to see how the proposed key frame choosing method can improve the performance of video shot retrieval.

TABLE I. DESCRIPTION OF THE TEST VIDEO SEQUENCES

Video	No. of Frames	No. of abrupt Cuts	No. of gradual changes
Earthquake Report	2000	6	9
Superman(Movie)	755	3	0
Friends (Movie)	2000	73	8
Tomb (Movie)	960	31	0
(Total)	5715	113	17

TABLE II. ABRUPT TRANSITION PERFORMANCE

Video	Histogram	Proposed
Earthquake Report	5	6
Superman (Movie)	3	3
Friends (Movie)	64	70
Tomb Raider (Movie)	28	25
	88.50%	92.04%

TABLE III. GRADUAL PERFORMANCE

Video	Histogram	Proposed
Earthquake Report	5	7
Superman (Movie)	0	0
Friends (Movie)	5	7
Tomb Raider (Movie)	0	0
	58.82%	82.35%

TABLE IV. FALSE ALARMS

Video	Histogram	Proposed
Earthquake Report	7	7
Superman (Movie)	0	0
Friends (Movie)	15	6
Tomb Raider (Movie)	22	6
	33.85%	14.62%

To evaluate the performance of the retrieval method, the Average Recall (AR) and Average Normalized Modified Retrieval Rank (ANMRR) [1] are used.

$$AR = \frac{1}{Q} \sum_{q=1}^Q \frac{nr(q)}{ng(q)} \quad (6)$$

$$ANMRR = \frac{1}{Q} \sum_{q=1}^Q NMRR(q) \quad (7)$$

$$NMRR(q) = \frac{MRR(q)}{K - \frac{ng(q)}{2} + 0.5} \quad (8)$$

$$MRR(q) = \sum_{i=1}^{ng(q)} \frac{r(i)}{ng(q)} - \frac{ng(q)}{2} - 0.5 \quad (9)$$

where  $ng(q)$  is the number of similar videos with query video  $q$ , there are  $Q$  query videos totally;  $nr(q)$  is the number of similar videos included in the first  $K$  results,  $K = \min\{4 \times ng(q), 2 \times GTM\}$ ,  $GTM = \max\{ng(q)\}$ ;  $r(i) = Rank$  if similar videos are all included in the first  $K$  results, otherwise  $r(i) = K + 1$ .

In the experiments, the proposed method is compared with the method TMOF in [1]. The experimental results are illustrated in Table V. It shows that our proposed method result in a lower ANMRR and a higher AR compared to TMOF. The reason is that more than one key frame are used to retrieval for the proposed method. The key frames have better representative power in video shot representation, and the video shots retrieval based on these key frames has better performance than TMOF.

TABLE V. PERFORMANCE OF DIFFERENT RETRIEVAL METHOD

	ANMRR	AR
TMOF	0.3635	0.7176
Proposed	0.2644	0.9256

## VI. CONCLUSIONS

In this paper, we present a new image sequence retrieval approach based on histogram difference of local image. Because frames with different contents have different interest points, which define different local images with different histogram, the proposed method can overcome the deficiency of traditional histogram-based method that different contents frames have similar histograms, and it can effectively detect gradual transitions for similar reasons. On the other hand, the large movement of camera will lead to false detection for Histogram-based method, while the proposed method is suitable for this situation, because when the camera moving with objects, most of interest points keep homologies in the regions of objects and the histogram of local images keep stable, thus it can distinguish between gradual transitions and camera motions effectively. The key frames selected by the proposed method have good representative power and can improve the performance of video shot retrieval than traditional method.

## REFERENCES

- [1] K. W. Sze, K. M. Lam and G. P. Qiu, "An optimal key frame representation for video shot retrieval," *In Proc. of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pp.270-273, Oct.2004.
- [2] A. Vadivel, M. Mohan, S. Sural and A. K. Majumda, "Object level frame comparison for video shot detection," *In Proc. of the IEEE Workshop on Motion and Video Computing*, pp.235-240, Jan. 2005.
- [3] S. W. Lee, Y. M. Kim and S. W. Choi, "Fast scene change detection using direct feature extraction from MPEG compressed videos," *IEEE Trans. Multimedia*, Vol.2, No.4, pp.240-254, Feb. 2000.
- [4] S. C. Pei and Y. Z. Chou, "Effective wipe detection in MPEG compressed video using macroblock type information," *IEEE Trans. Multimedia*, Vol.4, No.3, pp.309-319, Apr. 2002.
- [5] U. Gargi, R. Kasturi and S. H. Strayer, "Performance characterization of video shot-change detection methods," *IEEE Trans. Circuits and Systems for Video Technology*, Vol.CSVT-10, No.1, pp.1-13, 2000.
- [6] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," *In Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases*, Vol.3656, pp.290-301, Jan. 1998.
- [7] J. Zheng, F. M. Zou and M. Shi, "An efficient algorithm for video shot boundary detection," *In Proc. of International Symposium on Intelligent Multimedia, Video and Speech Processing*, pp.266-269, Oct. 2004.
- [8] C. Harris and M. Stephens, "A combined corner and edge detector," *Alvey Vision Conference*, pp.147-151, Sept 1988.
- [9] J. W. Rong, W. J. Jin and L. D. Wu, "Key frame extraction using inter-shot information," *Proc. of International Conference on Multimedia and Expo*, pp.571-574, Jun 2004.
- [10] K. S. Ntalianis and S. D. Kollias, "An Optimized Key-Frames Extraction Scheme Based on SVD and Correlation Minimization," *IEEE International Conference on Multimedia and Expo*, pp.792-795, Jul. 2005.
- [11] M. Mills, J. Cohen and Y. Y. Wong, "A magnifier tool for video data," *In Proc. ACM Computer Human Interface*, pp.93-98, May 1992.