Key frame extraction based on entropy difference and perceptual hash

Mi Zhang
School of Software Engineering
Xi'an JiaoTong University
Xi'an,China
mizhang@stu.xjtu.edu.cn

Lihua Tian
School of Software Engineering
Xi'an JiaoTong University
Xi'an,China
lhtian@mail.xjtu.edu.cn

Chen Li*
School of Software Engineering
Xi'an JiaoTong University
Xi'an,China
lylnnc@126.com

***Abstract*—Key frame extraction is a crucial step in content-based video retrieval. To accurately describe character of frames, various features like color, texture, shape can be integrated and used for key frame extraction. In this paper, we proposed an improved two-phase approach of key frame extraction based on entropy and perceptual hash. It weakens the threshold's direct influence on final results, and solves the problem of fading, sunlight and other information easily resulting in redundant key frames. Firstly, candidate key frames are selected with the use of golden-section partition and weighted histogram. Next, key frames are determined by the entropy values of candidate frames. Finally, a new method of perceptual hash is applied to remove redundant key frames. Experimental data set is created with videos from different domains like movie, cartoon, news etc. Results show that the proposed method is accurate and effective for key frame extraction. The selected key frames can be a good representative of main content.**

Keywords—block-weighted histogram; Golden-Section; Entropy; Perceptual hash

I. INTRODUCTION

The explosive growth of digital content requires the development of new technologies and methods to represent multimedia data. Keyframe extraction is an essential part in video analysis which provides an efficient summarization for video indexing and browsing. It reduces the amount of data required in video indexing and provides a framework for dealing with the video content. The selected key frames must summarize the characteristics of video, and the content of video can be tracked by all the key frames in time sequence.

Traditional algorithms are mostly based on structure of video and shot segmentation. Normally, a video is subdivided into a set of short segments or shots each of which contains similar content. Then, representative keyframes are selected from these segments by threshold. The weakness of this approach is that shot segmentation and threshold choosing directly impact the keyframe result set. To a great extent, the threshold limits the quantity and quality of key frames. In order to overcome the shortcomings of the algorithm above, reference[1] proposed an entropy-based method, which respectively uses entropy value as global feature and local feature to select key frames and remove redundant key frames. The results show that it works well when the image background is distinguishable, but when the gradient or color characteristics are similar between shots, it may miss key frames. Reference[2] put forward an approach based on image entropy and edge-matching rate. This method has made some improvements, but when the background of moving object is complex, there are errors in the edge features of the object which resulting in certain redundancy.

While analyzing the existing entropy-based keyframe extraction method, we proposed a novel algorithm based on color-entropy and perceptual hash.

II. RELATED WORK

A. Block-weighted Histogram on HSV Color Space

As a common color space, HSV is widely used in image processing. Compared with RGB, HSV is more intuitive to express the light, shade and bright degree of color[3]. Histogram is the statistical information of pixels, but it ignores the location of pixels[4]. As a result, images with similar color histogram sometimes have dramatically different appearances due to the distribution of pixels. To solve this problem, we proposed a histogram-based method with the preprocessing of Golden-section partitioning and weighting. Golden-section is a common skill used in photography, which divides a whole picture into nine rectangular grids of horizontal and vertical segmentation[7]. Shooting always follows a principle that main content should be distributed in the vicinity of golden section ratio or at the center of a picture. Therefore, we use the golden section method to identify the main objects and the relationship between objects and environment in the image.

B. Entropy Value of Image

Information entropy is first proposed by Shannon, which is a probability density function of random variables. For an image, entropy represents the amount of information contained in image. Suppose that $P(x_i)$ represents the pixel's ratio of x_i in the image. Image entropy is calculated as in (1) and (2):

$$p(x_i) = num(x_i) / (m \times n) \tag{1}$$

$$H = -\sum_{i=0}^{255} p(x_i) \times \log(p(x_i)) \qquad \sum_{i=0}^{255} p(x_i) = 1 \tag{2}$$

In formula (2), $0 \le p(x_i) \le 1$, x_i represents the gray value of pixel, m and n respectively represents the height and width of image. Num (x_i) means the number of pixel x_i in the image.

978-1-5386-2937-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ISM.2017.109

$P(x_i)$ is the probability of gray level x_i. When $p(x_i)=0$, the image information entropy makes no sense. Therefore, a constraint is added: if $p(x_i)=0$, $\log(p(x_i))=0$.

C. Perceptual hashing algorithm

Image perceptual hash technology is also referred to as digital fingerprints, which is a summary of multimedia information. Perceptual hash is a class of one way mappings from multimedia presentation to a perceptual hash value in terms of perceptual content. Each picture has a corresponding fingerprint. The closer the fingerprint value is, the higher the similarity of two pictures is. Implementation is as follows:

1) Reduce picture size: Remove details of pictures by ignoring the different size and proportion of pictures and leaving the basic information such as structure, shading, etc.

2) Simplify the color: Convert to a gray scale image to further simplify the calculation.

3) *Calculate DCT* : Calculate the DCT of each picture and obtain a 32*32 DCT coefficient matrix.

4) Reduce DCT : Keep the upper left corner of 8*8 matrix which shows the lowest frequency.

5) *Calculate average*: Calculate average gray value of all pixels.

6) Calculate Hash: Calculate hash value according to 8*8 DCT matrix. Then, compare the gray value of each pixel with the average. Pixels which is greater than or equal to the DCT mean are set to "1", otherwise set to "0". The string consisting of "0" and "1" above is the fingerprint of the image.

Perceptual hash algorithm avoids the effects of gamma correction and color histogram adjustment. In the meantime, it is not affected by the size of the image. As long as the structure of image keep unchanged, the hash value remains the same[8]

III. EXTRACTION ALGORITHM

The method proposed in this paper consists of three components: *A.* Candidate key frames are extracted by improved histogram with the use of golden-section partitioning and weighting. *B.* The candidate frame whose entropy value is a local maximum is selected as a key frame. *C.* Fingerprints are calculated by perceptual hash, and their Hamming distance is used as an indicator to evaluate the similarity of adjacent frames to eliminate redundant key frames.

A. Candidate Key Frame Extraction

The main content of video is usually located near the center or golden means in the image. According to Golden-Section principle, we divide each frame into three parts by golden-section and assign weight (w) for each block. Firstly, find the golden means (A, B, C, D) of each frame. Then continue to divide the frame into 6*6 rectangular blocks as is shown in Fig. 1. Weights fading from center to edge as it is shown in Fig. 2.

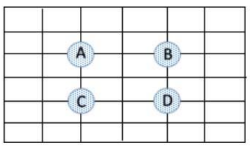


Fig. 1. blocked frame

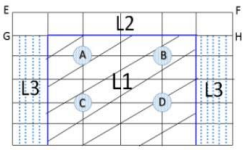


Fig. 2 weighted frame

- First part(L_1): main part labeled by notation ;
- Second part (L_2):rectangle of EFGH;
- Third part (L_3): the dotted block on the sides of frame .

1) Get total frame number (N) and sequence $F=\{f_1,f_2,...f_n\}$ of video. Subscript indicates the sequence number of frame in the video. Then, convert color space from RGB to HSV.

2)Frames are divided into sub-blocks as Fig. 2. Then calculate each block's color histogram. S_1, S_2, and S_3 represent the difference of corresponding blocks (L_1, L_2 and L_3) between frame f_i and f_{i+1} calculated by (3). Diff is the difference between two frames calculated by (4). The w_1, w_2 and w_3 are the weights of S_1, S_2 and S_3 respectively.

$$S(i,i+1)=\sqrt{\sum_j (H_{f_i,j}-H_{f_{i+1},j})^2}+\sqrt{\sum_j (S_{f_i,j}-S_{f_{i+1},j})^2}+\sqrt{\sum_j (V_{f_i,j}-V_{f_{i+1},j})^2} \quad (3)$$

$$diff(i,i+1)=\omega_1\times S_1+\omega_2\times S_2+\omega_3\times S_3 \quad (4)$$

3) Candidate key frame extraction: The threshold is calculated based on diff values of all adjacent frames. As is shown in (5).

$$Threshold=Mean(diff)+\alpha\times Square(diff) \quad (5)$$

Mean and Square are the average and square of diff. If diff(i,i+1)>threshold, frame numbered "i" is chosen to be a candidate key frame. Otherwise, compare frame "i" with next frame, until all frames are compared. "α" is a weighted factor, at this stage, we tend to set "α" lower as much as possible to ensure that key frames are not missing. Finally, candidate key frames' collection is obtained regarded as $K=\{k_1,k_2,...k_j\}$, and the subscript indicates the serial number of frames in the video.

B. Extract Key Frame by Entropy value

There is a strong correlation between two adjacent frames in the video sequence, and the changes between different shots present an order of magnitude difference. According to this law, we abandon the traditional method of selecting a global average or other coefficients of frames as threshold value. In this paper, the ratio of entropy difference is taken as the criterion of judgment. Specific steps are as follows:

Suppose k_i represents the candidate frame to be judged in collection K, f_i is the same frame in collection F. Entropy(k_i) represents the entropy value of the frame k_i, k_{cur} represents the selected key frame which is nearest to f_i in the video and has been selected by entropy value.

1) Calculate the entropy value of frame k_i by (1).

2) Select the prior candidate frame f_{i-1} and the next candidate frame f_{i+1} of f_i in the video. Then calculate Entropy(f_{i-1}) and Entropy(f_{i+1}) respectively.

3) Get key frame k_{cur} and calculate Entropy(k_{cur}). K_{cur} represents the key frame selected by entropy recently.

4) Calculate Entropy Diff: The difference of two frames' entropy is calculated by (6). Similarly, Entropy_diff(i,i+1) and Entropy_diff(i,cur) are calculated.

$$Entropy\_diff(i-1,i)=Entropy(k_i)-Entropy(k_{i-1}) \quad (6)$$

5) Calculate Diff_rate: Diff_rate is calculated as in (7) and (8). If these two entropy_diff are of different orders of

magnitude, the current frame is judged to be different from the previous key frame and selected as the next key frame.

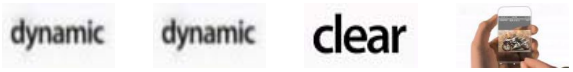
$$diff\_rate(i_1)=Entropy\_diff(i,cur)/Entropy\_diff(i-1, i) \quad (7)$$

$$diff\_rate(i_2)=Entropy\_diff(i,cur)/Entropy\_diff(i,i+1) \quad (8)$$

Candidate frame extraction mainly guarantees the integrity of video content. Then entropy value is used to optimize the candidate result and selected key frame. Above method avoids the direct effect of threshold on the result set. However, entropy-based algorithm is easily affected by several factors such as editing, fade and sunlight. In this case, the selected key frames based on entropy algorithm become unreliable as they may produce redundant key frames for the same shot[5].

C. Eliminate Redundant frame by Perceptual Hash

Key frames obtained from above step describe the main content completely, but there is a small amount of redundant frames requiring further processing. To eliminate redundant keyframes, segmented entropy technique is used in [1]. This method divides each frame into 64 segments and entropy of each segment is calculated individually. However, the performance of this method degrades when the video sequences contain transient changes and inserted graphics. Reference[2] removes redundant frames by edge matching rate. However, edge detection algorithm performs pixel by pixel, which is time-consuming. At the same time, it is sensitive to the brightness of frame. For example, Fig. 4. is a description of edge feature for Fig. 3. Though Fig. 3(a) and Fig. 3(b) seem to be the same, their edge contours are different which is due to the difference in brightness. As a result, both Fig. 3(a) and Fig.3(b) are selected as key frames by edge matching rate.



(a) (b) (c) (d)

Fig. 3. Test frames

(a) (b) (c) (d)

Fig. 4. Edge feature of frames in Fig.3

In our experiment, edge matching and perceptual hash algorithms are used to calculate the distance of images. The distance between frame x and y is defined as S(x,y). When S(x,y) equals 0, picture x is the same with picture y. Two similar pictures Fig. 3(a) and Fig. 3(b) and two pictures of great differences Fig. 3(c) and Fig. 3(d) are selected. The results are shown in TABLE I:

TABLE I. Edge_match and Phash_match

Comparison	Match_edge	Match_Phash
S(a,b)	0.00427	0
S(c,d)	0.00814	16

When the edge operator is used to calculate the distances of images, the values of S (a, b) and S (c, d) are very close, which means that the images' differences can not be accurately expressed. On the contrary, values obtained by our perceptual hash algorithm are S(a,b)=0 and S(c,d)=16, which can apparently distinguish the two situations. The steps of image comparison by perceptual hash is as follows:

1) Get the fingerprint of each key frame by perceptual hash;

2) Calculate Hamming distance $d(k_i,k_j)$ of adjacent frame. Then get average Hamming distance of all key frame;

3) Calculate the threshold: Threshold= β *average. If $d(k_i,k_j)$ < threshold, k_i is similar with k_j, then k_j is eliminated.

In conclusion, edge operator is sensitive to the brightness of image. Images with similar content may vary greatly due to local variations in brightness. Perceptual hash emphasizes the structure of picture and shows a stronger robustness for changes in picture size and the adjustment of color histogram.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Evaluation Criterion

Precision rate (P) and recall rate (R) are two criteria used to measure the validity and accuracy of extracted results. Recall rate reflects the comprehensiveness of results, which is used to measure the probability of missing key frames. These two criteria are mutually influenced with each other. The higher recall rate is, the lower precision rate will be. In order to get the appropriate measure, the harmonic mean value(F) of two rates is used. It is defined as follows:

$$P=\frac{N_c}{N_c+N_f}\times 100\% \quad (9) \qquad R=\frac{N_c}{N_c+N_m}\times 100\% \quad (10)$$

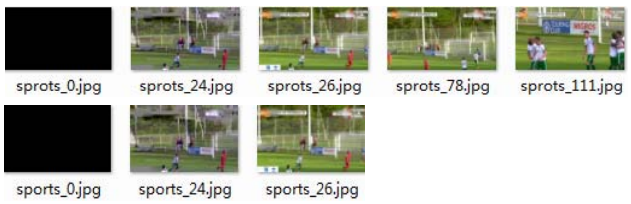
$$F=\frac{2\times R\times P}{R+P} \quad (11)$$

N_c, N_m and N_f respectively represent the number of correct key frames, the number of missing detection and the number of error detected key frames .

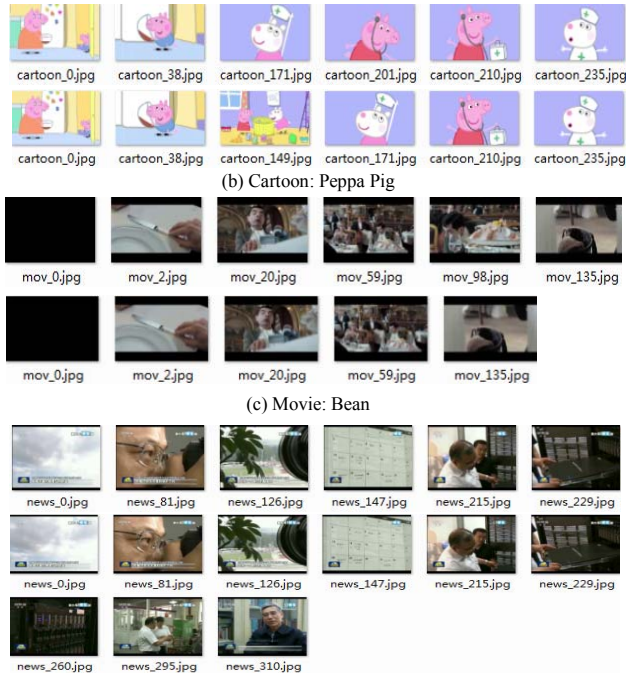
B. Results and Analysis

The algorithm was implemented on Visual Studio 2013 and OpenCV 3.0. The complexity of this algorithm is $O(n^2)$. In order to test the algorithm's effectiveness, thirty different videos which include news, cartoon, advertising, movies and sports are selected and tested for this algorithm. For each type of video, 6 videos are selected for testing. In the experiment, we make a comparison between golden-section blocked and non-blocked extraction. Data shows that when the weights of w_1, w_2, w_3 is 0.6, 0.3 and 0.1, there is a better recognition for the main content of frame. When the variance coefficient of threshold α =1.5, the redundant degree is low with fewer missing frames. When diff_rate>5 or diff_rate<0.2, the candidate frame should be selected as a key frame.

The results show that most values of F by blocked algorithm are greater than 90%, and only the value of sports video is relatively low and fluctuates near 85%. This is due to the rapid movement of objects. First, it is difficult to ensure the accuracy of standard key frames. Second, frames change quickly thus motion blur is caused. Fig. 5. respectively shows the key frame extracted by non-blocked and golden-section blocked algorithm.



(a) Sports : Football Competition



(b) Cartoon: Peppa Pig

(c) Movie: Bean

(d) News: CCTV

Fig. 5. Results by golden-section

The pictures in first row show the results of non-blocked method, and the second row is block-weighted key frame extraction result in Fig. 5. As what can be seen from above, there is a redundant frame cartoon_210.jpg in Fig. 5, and seven frames missed by non-blocked method. Almost no redundant frames show up in weighted-block result, in addition, Fig. 6. describes the contents of video in further details. From above, golden-section blocked algorithm can effectively highlight the main part of frames in the video. Compared with non-blocked algorithm, its result is more accurate. Furthermore, algorithm in this paper is compared with [1] which extracts key frame by local and global entropy. Refer to the results in TABLE II.

TABLE II. Compare with Reference[1]

video	R Ref.[1]	R Our method	P Ref.[1]	P Our method	F Ref. [1]	F Our method
mov0	97%	94%	100%	100%	98%	97%
mov1	86%	94%	94%	100%	89%	97%
mov2	91%	80%	72%	100%	80%	89%
news1	53%	100%	68%	100%	59%	100%
news2	91%	100%	87%	98%	89%	99%

As TABLE II shows, both of the two algorithms have a good performance for movie. In the meantime, their criterion value of F is comparable. But for News, method of [1] has a high redundant rate(P) and missing rate(R), and proposed algorithm is more efficient in News' key frame extraction. Fig. 6. compares the method in this paper with [1] and non-blocked algorithm. Results show that the algorithm in this paper is more stable and efficient than the other two algorithms. The accuracy and missing detection rate are both optimized in this paper.

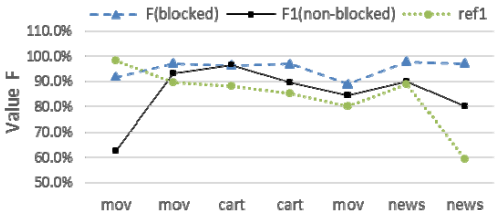


Fig. 6. Value F of three algorithms

In summary, the results indicate that key frames extracted by the new proposed algorithm can effectively depict the main content of video. Its results are obviously superior to the results of non-block extraction. Perceptual hash performs more accurate for similar images' recognition, more robust to image distortion, and simpler to compute.

V. CONCLUSION

The paper presents a novel key-frame extraction method based on entropy difference and perceptual hash. Two-phase key frame extraction weakens the direct influence of threshold on the result, and solves the problems of editing, fading, sunlight and other information which easily lead to redundancy. Experimental results show that this method has high accuracy and low miss rate. However, when the background of moving object is complex, there may be some redundant and missing frames, therefore further study is needed .

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 61403302 and the Fundamental Research Funds for the Central Universities No.XJJ2016029.

REFERENCES

[1] SP Algur ,R Vivek.Video Key Frame Extraction using Entropy value as global and local feature[J].Computer Vision and Pattern Recognition,2016

[2] L Ren,Z Qu.Key frame extraction based on information entropy and edge matching rate[J] International Conference on Future Computer & Communication,2010,3:V3-91-V3-94

[3] L.C.Jiang,G.Q.Shen,G.X.Zhang.An image retrieval algorithm based on HSV color segment histograms[J]. Mechanicaland Electrical Engineering Magazine,2009,26(11),pp.54-60.

[4] H.Y.Liu ,T.Li.Key Frame Extraction Algorithm Based on Improved Block Color Features and Second Extraction[J] Computer Science, 2015,42(12):pp.307-311.

[5] PM Kamde,S Shiravale.Entropy Supported Video Indexing for Content based Video Retrieval[J].International Journal of Computer Applications,2013, 62(17):734-734

[6] H.Y.Liu ,T.Li.Key Frame Extraction Algorithm Based on Improved Block Color Features and Second Extraction[J] Computer Science, 2015,42(12):pp.307-311.

[7] D.B.Hu. The shooting technique of still and moving picture in TV [J] Journal of Shenyang Normal University , 2002, 20(4),pp.275-277.

[8] Rudakov,Vasiutovich. Analysis of Perceptual Image Hash Functions[J] LCC:Computer engineering. Vol 0, Iss8, Pp 269-280 (2015)