Key Frame Extraction Based on Binocular Cameras

Zili Li, Luoxian Zhang and Haiying Xia

ABSTRACT

Key frames provide the effective summary information for video retrieval and browsing, which mostly represents the main content for video sequences. The statistical results prove that color, as a feature that represents image content, proved widely used in the key frame extraction. Stereo correspondence has been an indispensable research direction in the field of computational stereo vision. Stereo vision is introduced for the purpose of improving previous key frame extraction algorithms, which is the key contribution of our work presented in this paper. A novel image similarity measurement is constructed by combining the color moment with SAD. Experimental results demonstrate that the average evaluation coefficient obtained by the improved algorithm in this paper reach up to 95% in the situation of camera stillness and motion. The novel algorithm solves the issue of redundancy which stems from the large changes of content among consecutive frames in the process of extracting key frames using monocular cameras, and performs more robust compared with traditional algorithms.

KEYWORDS

Key Frames, Stereo Correspondence, Sum of Absolute Difference(SAD), Color Moment, Similarity Measurement.

INTRODUCTION

The video shot for the same scene often has a large amount of redundant information, and only a small scale of frames, key frames, can be used to represent the main content of a video sequence. Key frames are useful in a wide variety of applications like summarizing, storing, indexing, retrieving video clips[1] etc. The use of key-frames greatly reduces the amount of data required in video indexing and provides an organizational framework for dealing with video content[2]. Feature vector representation and similarity measurement are very critical for the retrieval performance of a CBIR system[3]. Therefore, the main work is focused on the two points above in relation to key frame extraction. An enormous amount of scholars have made great efforts towards methods of extracting key frames for the purpose of laying the roots for summarizing a video effectively and succinctly. After the video stream being segmented into shots, a natural and simple key frame extraction method is to regard the first frame of each shot as the key frame of the shot[4]. Wolf et al.[5] proposed a selection of key frames using an algorithm based on optical flow analysis.

Zili Li, Luoxian Zhang, Haiying Xia
School of Electronic Engineering, Guangxi Normal University, Guilin 541004, China
Zhang et. al.[6] presented to use content-based criteria to extract key frames. However, the algorithms above are only concentrated on the extraction of key frames based on consecutive frames of video sequence, and further research is needed to extract key frames based on binocular cameras. Consequently, stereo vision is introduced in this paper for the purpose of improving of previous key frame extraction algorithms. Stereo vision has always been the hot spot of computer vision research, which simulates the stereoscopic perception process of human vision to the greatest extent. Stereo correspondence is an indispensable research direction in the field of computational stereo vision.

Having been widely used in many stereo vision systems, block matching[18] has low computational complexity and many acceleration methods so that it proves to be capable of implementing hardware easily[7,8,9]. In 2003, Brown [18] proposed quite a few widely used measure functions: sum of squared differences (SSD), sum of absolute differences, normalized cross correlation (NCC), etc. Rhemann et al.[10] introduced a smoothing filtering technique with edge protection into the local stereo correspondence algorithm proposed in 2011, which was the optimal local matching algorithm at that time. Tao et al.[11] improved the fast stereo correspondence algorithm with SAD as matching criterion, so that the computational cost of the method was reduced by more than 55%. Zhang et al.[12] proposed a key frame extraction approach based on entropy difference as well as perceptual hash, which effectively eliminates redundant frames in video sequence. Higher order color moments based on block histogram were put into use by Poonam S. Jadhav et al.[13] to decompose the shot, and then the frames with the highest mean and standard deviation from each shot are selected as key frames. Dhagdi and Deshmukh[14] proposed a key frame extraction method based on histogram difference and edge matching rate. Lu et al.[15] represented each frame with two tuple features, that is, entropy in color and occupancy rate.

Previous studies show that composite descriptors in combination with different image features are proved necessary in key frame extractions. In this paper, we propose stereo vision and construct a composite descriptor that combine color moment with SAD.

THEORETICAL PRINCIPLE

COLOR MOMENT

The HSV color space is developed to provide an intuitive representation of color and to imitate the way in which humans perceive and manipulate color[16]. Therefore, all algorithm tests in this paper are performed by consideration of HSV space. Color moment[17] is a close feature describing color distributions, and it applies the concept of moments in the linear algebra to the description of image features. The first three moments in each color channel are selected, which were denoted as follows:

\[
\mu_i = \frac{1}{N} \sum_{j=1}^{N} P_{i,j} \quad (1)
\]

\[
\sigma_i = \left( \frac{1}{N} \sum_{j=1}^{N} (P_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}} \quad (2)
\]

\[
S_i = \left( \frac{1}{N} \sum_{j=1}^{N} (P_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}} \quad (3)
\]
where \( P_{i,j} \) represents the probability of pixels with gray level \( j \) in the \( i \)th color channel component and \( N \) represents the number of pixels in a color image. Considering that each pixel has three channels in the color space, the three color moments only need 9 components to describe an image. If \( H \) and \( I \) represent the color distribution of the front and rear frames respectively, an equation is constructed for similarity measures as follows:

\[
d_{\text{mom}}(H, I) = \sum_{i=1}^{r} w_{i1}|E_{1}^{i} - E_{1}^{r}| + w_{i2}|\sigma_{1}^{i} - \sigma_{1}^{r}| + w_{i3}|S_{1}^{i} - S_{1}^{r}|
\]

(4)

where \( i \) and \( r \) represent color channel and the number of total channels, \( E_{1}^{i} \) and \( E_{1}^{r} \) are the first moment (mean), \( \sigma_{1}^{i} \) and \( \sigma_{1}^{r} \) are the second moment (variance), \( S_{1}^{i} \) and \( S_{1}^{r} \) are the third moment (skewness) of the image, \( w_{i1}, w_{i2} \) and \( w_{i3} \) are user-defined weights. The retrieval efficiency of color moment method proves low, hence it is often used in combination with other image features to narrow down the retrieval range.

**SAD**

The basic principle of stereo vision can be depicted to observe the same scene from two (or more) viewpoints so as to obtain perceptual images from heterogeneous perspectives. The 3D information of the scene is attained by calculating the position deviation—disparity between the pixels of the image by triangulation principle. Brown[18] divided stereo correspondence algorithms into two categories. Local correspondence methods are exclusively restricted from the localized area around the pixel, these restrictions include algorithms based on blocking matching, feature matching as well as gradient methods. The algorithm above proves fast but fail to keep accurate and mostly applied for preliminary screening before multistage processing. SAD (sum of absolute difference) is a concise and fast region-based image matching algorithm, of which the basic idea is to sum the absolute difference between the corresponding values of each pixel, and then the measured similarity between two image blocks is further evaluated according to SAD:

\[
C(u, v, d) = \sum_{(i,j) \in W} |I_l(u + i, v + j) - I_r(u + i + d, v + j)|
\]

(5)

where \( I_l \) and \( I_r \) are the gray values of pixels of the left and right image respectively, \( W \) represents the neighborhood window centered on a point in the image, \( C(u,v,d) \) is the matching cost after accumulative calculation.

**THE CONSTRUCTION OF WEIGHTED SYNTHESIS ALGORITHM**

The fact that objects with cameras are usually in motion indicates that even in ideal state the subtle change in the relative position of the cameras seems inevitable in a scene where multiple cameras are required. The content of successive frames of video sequence always changes greatly in the case of fast motions. Consequently,
traditional key frame extraction algorithms based on consecutive frames easily give rise to redundancy. In this paper, a novel key frame extraction algorithm is proposed. There are two methodological innovations: (1) The difference between left and right frames of the binocular cameras at the same time are compared rather than successive frames of video sequence, and the current frame is selected as key frames, only if the similarity is less than the threshold (that is, when the jitter is severe.) (2) Color moment[17] and SAD[18] are integrated into a single equation in weighted synthesis, and a distance of the absolute value with weights is chosen as an indicator to evaluate image similarity:

\[ d(L, R) = \sum_{i} w_i \left| E_i - E_R \right| + \sum_{i} \left| I_L (u + m, v + n) - I_R (u + m + d, v + n) \right| + \sum_{i} \left| \sigma_i - \sigma_R \right| + \sum_{i} |S_i + S_R| \]

where \( d(L, R) \) represents the similarity of left and right frames, \( W \) represents the neighborhood window centered on a point in the image, \( i \) is the channel of color, \( E \), \( \sigma \) and \( S \) are the first, second and third moment of the image respectively, \( I_L \) and \( I_R \) represent gray value of pixels, \( w_{11} - w_{44} \) are user-defined weights. The new algorithm uses 10 components to describe features in an image.

**EXPERIMENTAL ANALYSIS**

**DATA ACQUISITION**

The new algorithm is tested in Visual Studio 2013 and OpenCV 3.4.1. In order to verify that the algorithm has good performance under the condition of camera jitter, we let the car drive on the normal road and record the videos. The specific implementation is as follows: First, two cameras which have the same type are fixed to the left and right sides of the car respectively in the same horizontal line. Second, adjusting the cameras’ angle to ensure that the overlapping field of view ranging from 30 to 60 percent. The last step, synchronizing the videos recorded by the two cameras. The frame rate of all the cameras set in this paper is 30fps. As an empirical value, the 7 × 7 criterion is put to use as the window size of SAD.

**PERFORMANCE EVALUATION**

In comparison with the key frame extraction algorithm based on SAD, the improved algorithm presents a new strategy—color moment in the process of measuring similarity, thus the number of key frames extracted remains less and the accuracy is improved. Although the latter increases the computational cost of color moment accordingly, efficiency should not be at the expense of correctness. Theoretically, the proposed algorithm in this paper combines the difference between color distribution and gray level, and demonstrates better screening performance for binocular-based key frame extraction.

Moreover, the improved method proves to be a relatively concise approach. What keeps generally agreed is that when the the car runs faster, the cameras wobbles more
violently; when the car runs slower, the jitter is more smooth. The speed of the car is set to the unique variable in the implementation, and the test cases consist of multiple videos in the same scene. Figure 1. shows a part of key frames extracted by the new algorithm at different speeds.

![Figure 1. Key frames extracted at different speeds. (a) v=10km/h. (b) v=20km/h. (c) v=30km/h.]

As shown in Figure 1, the pictures on the first and second line respectively represent key frames extracted by cameras in left and right at the same time. Figure 1(a) has no redundant frames, Figure 1(b) has less redundant frames and Figure 1(c) has abundant redundant frames. The pictures above show that with the increasing of speed, the number of extracted key frames as well as redundant frames increase accordingly. Precision rate (P) and recall rate (R) measure the effectiveness and accuracy of the proposed algorithm. Precision rate (P) reflects the proportion of target results in the captured results; Recall rate (R), as the name implies, reflects the proportion of target categories recalled from concerned sample. P is proportional to R under ideal situation, while these two indices always performs contradictory. The harmonic mean (F) which proves feasible in reference to evaluate the performance of the new algorithm is defined as follows:

\[ P = \frac{A}{A+B} \]  \hspace{2cm}  \( R = \frac{A}{A+C} \)  \hspace{2cm}  \[ F = \frac{2 \times R \times P}{R+P} \]
where A, B and C respectively represent the correct key frames, the wrong key frames and the number of missed key frames extracted. The evaluation coefficient F ranges from 0 to 1, and high value of F illustrates excellent performance of an algorithm.

| Speed (km/h) | P  | R  | F  | Total Frames |
|--------------|----|----|----|--------------|
| v=10         | 0.94 | 0.98 | 0.96 |              |
| v=20         | 0.91 | 0.85 | 0.88 |              |
| v=30         | 0.92 | 0.89 | 0.90 | 317          |
| v=40         | 0.84 | 0.89 | 0.86 |              |
| v=50         | 0.53 | 0.75 | 0.62 |              |

As shown in Table I, the effectiveness of the improved algorithm in this paper is directly associated with the speed of the camera’s motion. The camera moves comparatively stable when v=10km/h, and the number of extracted key frames remains the least and the accuracy gets the highest. When v >= 20km/h, F is in inverse proportion to the speed and reaches its peak at v = 30km/h. When v >= 50km/h, the jitter of cameras is too severe, which has little referred impact for the performance evaluation of the new algorithm.

**DATA ANALYSIS**

Furthermore, the novel algorithm is compared with the other three algorithms to verify whether it has the advantages of time consuming and accuracy. The other three algorithms are: algorithm solely based on SAD, method in[19] and[12]. Singh V P and Srivastava R[19] proposed a simple and effective image retrieval approach using color-invariant moments. Videos of test sets are obtained by cameras under moving condition and static condition respectively in order to assess the robustness of the proposed algorithm in different scenarios. Video 1 and Video 2 are two sports videos recorded by cameras under static condition. Video 3 and Video 4 are videos obtained by cameras at the speed of 30km/.

| Video | Total Frames | F SAD | F Ref.[19] | F Ref.[12] | F Proposed |
|-------|--------------|-------|------------|------------|------------|
| 1     | 2000         | 0.58  | 0.67       | 0.97       | 0.97       |
| 2     | 4000         | 0.60  | 0.69       | 0.97       | 0.98       |
| 3     | 2000         | 0.43  | 0.47       | 0.69       | 0.91       |
| 4     | 4000         | 0.50  | 0.53       | 0.73       | 0.94       |
| Average of F | 0.5275 | 0.59 | 0.84 | 0.95 |

As shown in Table II, the proposed algorithm achieves optimal results both under moving and static conditions. In addition, the F values of the proposed algorithm in the two states appear extraordinary close to each other, both of which are above 90%. The experimental result of algorithm in[12] seems almost the same as the proposed algorithm in the state of still. However, the proposed algorithm outperforms the other...
three algorithms in the state of motion, which indicates the fact that the new algorithm based on binocular cameras remains outstanding robust against camera jitter. Fig.2 shows that the improved key frame extraction algorithm based on binocular cameras is optimal compared with monocular methods referred above.

![Comparision between monocular and binocular cameras.](image)

**CONCLUSION**

Stereo vision is introduced for the purpose of the improvement of previous key frame extraction algorithms in this paper, and a novel image similarity measure is constructed by combining color moment with SAD. The experimental results show that the proposed algorithm achieves high evaluation coefficient and keeps robustness in the situation of stillness or in motion status. Moreover, the extracted key frames are able to represent the video content very well. The new algorithm solves the issue of redundancy which stems from the large changes of content among consecutive frames in the process of extracting key frames using monocular cameras, and has a high value in the key frame extraction on mobile platform.

**ACKNOWLEDGMENTS**

This work is supported by the National Natural Science Foundation of China (No.61762014).
REFERENCES

1. B. R. Mounika, O. Prakash and A. Khare, "Key Frame Extraction using Uniform Local Binary Pattern," 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), Allahabad, India, 2018, pp. 87-91.
2. A. Nagasaka, Y. Tanaka, "Automatic video indexing and full-video search for object appearances" in Visual Database Systems II, Amsterdam: Elsevier, pp. 113-127, 1992.
3. V. P. Singh and R. Srivastava, "Improved image retrieval using color-invariant moments," 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, 2017, pp. 1-6.
4. Yingying, Zhu, and Z. Dongru. "An Approach of Key Frame Extraction Based on Video Clustering." Computer Engineering (2004).
5. W. Wolf, "Key frame selection by motion analysis," 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 1996, pp. 1228-1231 vol. 2,doi: 10.1109/ICASSP.1996.543588.
6. Zhang, Hong Jiang, et al. "An integrated system for content-based video retrieval and browsing." Pattern Recognition30. 4(1997):643-658.
7. P. Corke and P. Dunn, "Real-time stereopsis using FPGAs," TENCON '97 Brisbane - Australia. Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (Cat. No.97CH36162), Brisbane, Queensland, Australia, 1997, pp. 235-238 vol. 1.
8. K. Konolige, "Small Vision Systems: Hardware and Implementation", Proc. Int'l Symp. Robotics Research, 1997.
9. O. Veksler, "Fast variable window for stereo correspondence using integral images," 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., Madison, WI, USA, 2003, pp. I-I.
10. C. Rhemann, A. Hosni, M. Bleyer, C. Rother and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 3017-3024.
11. Tangfei Tao, Ja Choon Koo and Hyouk Ryeol Choi, "A fast block matching algorithm for stereo correspondence," 2008 IEEE Conference on Cybernetics and Intelligent Systems, Chengdu, 2008, pp. 38-41.
12. Zhang M, Tian L, Li C.,"Key Frame Extraction Based on Entropy Difference and Perceptual Hash," 2017 IEEE International Symposium on Multimedia, 2017, pp.557-560.
13. Mrs. Poonam S. Jadhav, Dipi S. Jadhav, Video Summarization Using Higher Order Color Moments (VSUHCM),Procedia Computer Science, Volume 45,2015, pp. 275-281.
14. Dthagdi, Mr. Sandip T., and P. R. Deshmukh, "Keyframe Based Video Summarization Using Automatic Threshold & Edge Matching Rate."
15. Lu, Guoliang, et al. "Unsupervised, efficient and scalable key-frame selection for automatic summarization of surveillance videos," Multimedia Tools and Applications, 76(5), 2017, pp. 6309-6331.
16. B. S. Manjunath, J., Ohm, V. V. Vasudevan and A. Yamada, "Color and texture descriptors," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 703-715, June 2001.
17. Stricker M A, “Orengo M. Similarity of Color Images,” Proceedings of SPIE—The International Society for Optical Engineering, 1995, pp. 381-392.
18. M. Z. Brown, D. Burschka and G. D. Hager, "Advances in computational stereo," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 8, pp. 993-1008, Aug. 2003.
19. V. P. Singh and R. Srivastava, "Improved image retrieval using color-invariant moments," 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, 2017, pp. 1-6,doi: 10.1109/CIACT.2017.7977378.